ASPECT EXTRACTION FOR SENTIMENT ANALYSIS IN ARABIC DIALECT

by

Alawya Adnan Alawami

BS Computer Science, King Saud University, 2006

MS Information Science, University of Pittsburgh, 2010

Submitted to the Graduate Faculty of

School of Information Science in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2016

UNIVERSITY OF PITTSBURGH

SCHOOL OF INFORMATION SCIENCE

This dissertation was presented

by

Alawya Adnan Alawami

It was defended on

May 30, 2017

and approved by

Dr. Stephen Hirtle PhD, Professor, School of Information Science, University of Pittsburgh

Dr. Yu-Ru Lin PhD, Assistant Professor, School of Information Science, University of Pittsburgh

Dr. Mona Diab PhD, Associate Professor, School of Computer Science, George Washington University

Dissertation Advisor: Dr. Michael Spring PhD, Associate Professor, School of Information Science

ASPECT EXTRACTION FOR SENTIMENT ANALYSIS IN ARABIC DIALECT

Alawya Alawami, M.S

University of Pittsburgh, 2017

The increase of the user-generated content on the web led to the explosion of opinionated text which facilitated opinion mining research. Despite the popularity of this research field on English text and the large number of Arabic speakers who contribute continuously to the web content, Arabic opinion mining has not received much attention due to the lack of reliable NLP tools and an accepted/comprehensive dataset. While English opinion mining has been studied extensively, Arabic opinion mining has not received as much attention. The work that exists in sentiment analysis is limited to news, blogs written in Modern Standard Arabic and few studies on social media and web reviews written in Arabic dialect. Moreover, most of the work done has been done at the document and sentence level and to best of our knowledge, there is no work on a more fine-grained level. In this work, we take a more fine-grained approach to Arabic opinion mining at the aspect level through experimentation with methods that have been used in English Aspect extraction. Further, we are also contributing a dataset that can be used for further research on Arabic dialect.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENT

First and foremost, I would like to express my deepest appreciation to my advisor Dr. Michael Spring for his continued support, patience and motivation. This dissertation would not have been possible without him.

My sincere thanks also go to Dr. Mona Diab for her guidance and precious time for supporting this research. I am also grateful for her continued work in Arabic NLP which was the first light to this dissertation.

I would like to thank the rest of my thesis committee– Dr. Yu-Ru Lin, and Dr. Stephen Hirtle for their insightful comments and guidance.

I would also like to acknowledge and appreciate Dr. Nizar Habash and his team for their contribution of Arabic dependency parser which was very valuable to continue this research. I would also like to acknowledge Dr. Xin Zhao for providing the code for JST and patiently taking the time to communicate with me.

I also thank my colleagues William Garrard, Jonathan Grady, Jessica Benner and Fatimah Radwan for their helpful ideas, opinions, and discussion throughout my PhD studies.

Last but not least, there are no proper words to convey my deep gratitude for my husband Mohammed, my parents and the rest of my family for their continuous support and encouragement during my studies. 1

1. INTRODUCTION

Sentiment Analysis, also called opinion mining, is “the field of study that analyzes people’s opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes” (Liu, 2012). It is the step that follows subjectivity analysis and aims at classifying subjective phrases to positive, negative, natural or mixed.

Sentiment analysis can be done on three levels: document, sentence, and aspect level. Document level sentiment analysis assumes that each document holds opinions about one entity. Thus, sentiment classification at this level classifies the overall opinion about the entity. Sentence level goes a little deeper to classify the sentiment of each sentence that is later used to classify the overall sentiment of the document. A more fine-grained analysis is done at the aspect level where the system finds what the author likes or dislikes about the entity. It is also called feature based or attribute based sentiment analysis.

The goal of aspect level sentiment classification is to identify aspects along with the sentiment expressed on them. For example: “the food is delicious but the service is very slow” states the opinion of the reviewer on two aspects: food and service. Food sentiment is positive while service sentiment is negative. In document or sentence level sentiment analysis, the system does not discover this difference and the positive sentiment will offset the negative sentiment

leading to an overall positive or negative sentiment of the reviewer regardless of diversity in their views about aspects.

Aspect level sentiment analysis involves many tasks such as aspect extraction, aspect sentiment classification, aspect categorization. The focus of this research is on the first step of identifying aspects and extracting them from a review as well as a preliminary step of identifying opinion words related to the extracted aspects.

1.1 CONTRIBUTION

This work contributes to the field of Arabic sentiment analysis by identifying and assessing the performance of the methods that can be used to extract aspect terms and the opinion words related to them. For the purpose of extracting aspect terms, a corpus was built and will be made available to other academic researchers in the hope of moving the field forward. The corpus will also facilitate research on sentiment analysis at the aspect level for Arabic dialects. This research is limited to Arabic but the approach and methods can be applied to other low resource languages.

1.2 STATEMENT OF THE PROBLEM

Arabic is a native language for 290 million people around the world. Despite this, the Arabic Natural Language Processing field has a limited number of tools due to many challenges related to the nature and usage of the language. Unlike English, Arabic opinion mining did not receive much attention and to the best of our knowledge has not been explored at all possible levels. In

response to this problem, this research aims at investigating the methods that can be used for aspect level sentiment analysis including the methods used for extracting aspect using limited resources.

Sentiment analysis can be defined formally as follows. Given a review that consists of one or more sentences, the aim of sentiment analysis is to extract every sentence that contains aspect and classify its sentiment. This can be represented by the triple (E_i, A_i, S_i) where

E: the entity being reviewed,

A: represents implicit aspect

S: represents sentiment expressed on the aspect.

In the case of reviews, the opinion holder is omitted from this relation. The opinion holder is the person who reviewed the entity.

1.3 PURPOSE OF THE STUDY

The rise of the social media has generated a considerable amount of opinionated text. This opinionated text is valuable to individuals and companies in the process of making decisions. Given the vast amount of this information organizations as well as the public are in need of a system that can summarize those distributed opinions. This study focused on studying the methods that can be used to extract aspects from reviews written in Arabic dialect. More specifically, we worked with Arabic reviews by evaluating the performance of the existing techniques for aspect extraction that have been used for other languages, identifying the best performing approach and applying it for the purpose of sentiment analysis.

1.4 RESEARCH QUESTIONS

There are five questions to be answered by this research:

- What are the methods that can be used for Aspect Extraction?
- Which of these methods can be applied to Arabic reviews giving the limitation of NLP tools available?
- Can translation and English NLP tools be used to facilitate aspect extraction?
- How do these methods compare?
- Which method perform best and thus can be used for the purpose of sentiment analysis at the aspect level?

1.5 DEFINITION OF TERMS

Aspects: that are part of or related to an entity. In the case of restaurant reviews, menu, service, food, atmosphere can all be aspect terms. They can be explicit (the food is really good) or implicit (the restaurant is very expensive "aspect: price"). They are also referred to as features or attributes.

Aspect Extraction Task: (also called opinion target identification): Identifying words or phrases that are considered attributes of an entity being reviewed.

Conditional Random Field (CRF): Supervised statistical modeling method that is used for pattern recognition and machine learning.

Dialect Arabic (DA): The spoken form of Arabic language. It varies widely from region to region and the written form of it is widely used in social media.

Entity: the thing that is being reviewed.

Hidden Markov Model (HMM): Model in which the system being modeled follow a Markov process with unseen states. This model has a wide application in part of speech tagging, speech recognition, bioinformatics and many others.

Latent Dirichlet Allocation (LDA): Model that assume a collection of documents is a mixture of topics and represent each document as a set of topic probabilities.

Machine Translation (MT): Software to translate text or speech from one language to another.

Modern Standard Arabic (MSA): The standard form of Arabic which is recognizable by Arabic speakers, taught in school and used in newspapers and books but not spoken daily and in informal settings.

Natural Language Processing (NLP): A field of computer science that is concerned with allowing machines to derive meaning from natural form of human languages through developing a wide of variety of tools to aid in this task.

Sentiment: (also called polarity): attitude, feeling, or opinion toward an entity.

Sentiment analysis: using text analysis techniques to identify the terms or phrases that define the sentiments of a document, sentence or phrase and classifying them as positive, negative and natural.

Subjective Analysis: Differentiating objective phrases “the final game is on Sunday” from subjective phrases “the final game was awesome!”.

Subjective sentences: Sentences that holds opinion, emotions, or attitude.

Supervised Methods: Machine learning methods that rely on labeled training data to infer a model or a function that can be utilized for unseen data.

Objective sentences: sentences that hold facts and do not contain opinions

Part of Speech Tagging: Assigning part of speech tags to word of a corpus in their context.

Unsupervised Methods: Machine learning methods that do not rely on training data and try to find a hidden structure (model) in unlabeled data.

2.0 LITREATURE REVIEW

Given the focus on opinion mining for Arabic dialect, the literature review can be divided in two parts. The work that has been done in English for aspect extraction (section 2.1) and the work that has been done for opinion mining for Arabic dialect (section 2.2).

2.1 ASPECT EXTRACTION FOR THE TASK OF SENTIMENT ANALYSIS

Aspect based sentiment analysis refers to a sub field of sentiment analysis that recognizes each phrase that contains sentiment and extracts the aspect to which they refer. In this task, the system looks for aspect related to the entity being discussed. In general, reviews have two kinds of opinions: “GENERAL” about the whole entity such as “I like it” and SPECIFIC about certain attributes as in “The food is great”. This proposal deals with identifying specific opinion sentences since they usually hold aspects that the opinion is expressed on. There are many methods that have been used to extract aspects in reviews in English language; they can be divided into three approaches: supervised, semi-supervised and unsupervised methods.

2.1.1 Supervised Aspect Extraction Methods

This approach treats the problem of extracting aspect as an information extraction task. It relies on training data, which is manually labeling aspect and sentiment terms in the dataset. The most dominant approach in this field is sequential labeling techniques, such as Hidden Markov Chain (Rabiner & Juang, 1986) and Conditional Random Fields (Lafferty, McCallum, & Pereira, 2001). Jin, Ho, and Srihari (2009) employed a lexicalized HMM model to learn patterns that are used in extracting aspects and opinion. Jakob and Gurevych (2010) used CRF in multiple domains in an attempt to overcome the problem of domain dependency. They incorporated domain independent features such as part of speech tags, word distance, syntactic features and opinion sentences. Li et al. (2010) used Skip chain-CRF and Tree –CRF and skip-tree CRF to jointly extract opinion and aspects. CRF was also used by (Choi, Cardie, Riloff, & Patwardhan, 2005).

2.1.2 Semi-supervised Aspect Extraction Methods

The semi-supervised method used in aspect extraction employs opinion and target relation. This approach makes use of the idea that every sentiment word that exists in the document belongs to the nearest noun or noun phrase. This technique was developed by Hu and Liu (2004a). A similar approach was used by Zhuang, Jing, and Zhu (2006); it uses dependency trees to identify this relationship. Unfortunately, both approaches require the use of good Part-Of-Speech tagger and a good parser and such tools do not exist yet for Arabic dialect. We plan to experiment with this approach through the use of machine translation along with an English POS tagger and Parser.

2.1.3 Unsupervised Aspect Extraction Methods

There are many aspect extraction methods that can be classified as unsupervised methods. One line of research incorporates NLP tools to assist in the extraction. Most of this work is based on the observation that users who reviewed the same entity use a vocabulary that converges(Hu & Liu, 2004a). Since aspect terms are nouns or noun phrases, the part of speech tagger is used to find nouns and noun phrases. Then the most frequent nouns are kept based on a threshold determined experimentally. The earliest work on aspect extraction is based on this method, Hu and Liu (2004a) used frequency along with association mining and some pruning strategies to find aspect terms. This approached was improved through using PMI score to get rid of noun phrase that are not aspect terms. This approach is challenging to implement with the Arabic dialect because of the lack of reliable NLP.

The other line of aspect extraction research using unsupervised methods is based on building topic models. It is used to discover topics in large collections of text. Topic modeling is a generative model which assumes that each document in the collection is a mixture of topics and each topic is a probability distribution over words. The result of the method is a set of words along with their probabilities, each set represents a topic. Topic modeling has been used to extract topics from large collections of text and similarly has been applied to extract aspect terms from reviews.

There are two main topic modeling techniques: Probabilistic Latent Semantic Analysis (pLSA) (Hofmann, 1999) and Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003). Mei, Ling, Wondra, Su, and Zhai (2007) built the aspect sentiment mixture model based on pLSA. This mixture model consists of three models: aspect model, positive model and negative model. Most

of the other proposed models are based on LDA (Brody & Elhadad, 2010; C. Lin & He, 2009; Titov & McDonald, 2008a). Topic models are explored in more detail in section 5.7.

Topic modeling has many limitations. First, it needs a large volume of data and heavy tuning to achieve good results. Many of those models use Gibbs sampling which produces different results each run. Consequently, researchers spend a significant amount of time in parameter tuning. Finally, topic modeling is very useful in finding global aspects in a dataset but may fail in finding the most frequent local aspect which is more relevant to the entity being reviewed. On the other hand, it has the advantage of extracting aspect categories instead of extracting aspects independently and going through the process of categorizing them. It has the advantage of combining aspect extraction with finding their sentiment.

2.2 ARABIC OPINION MINING RESEARCH

Arabic opinion mining research has started to gain some attention recently because of it is potential in gaining more insight into users' opinion. The focus of this research is on the Arabic opinion mining.

2.2.1 Arabic Corpora

The successful progress of sentiment analysis systems depends largely on the availability of annotated corpora that can be used in training and testing. Unfortunately, the Arabic language lacks NLP resources and publicly available corpora so most researchers build their own datasets. To date and to the best of our knowledge, there are a limited number of Arabic corpuses available

for sentiment analysis and none of them were developed for the aspect extraction task. Korayem, Crandall, and Abdul-Mageed (2012) compiled an extensive survey of the current Arabic Corpora available for sentiment analysis at the document and sentence level.

2.2.2 Arabic Sentiment Analysis

Sentiment analysis resembles text classification which depends on finding a topic for the document (ex: sport, news, movies etc.). Sentiment analysis is concerned with classifying subjective phrases into positive, negative, natural and mixed categories. Based on that, text classification approaches are the main tools in analyzing sentiments. There are different approaches that were used in the literature to analyze English sentiment at different levels but most of the work for Arabic sentiment analysis is limited to the document and sentence level as discussed in the following section. To the best of our knowledge, there has been no work at the aspect level that we establish the basis for it in this work.

2.2.2.1 Document level Arabic Sentiment Analysis: The document sentiment analysis task can be considered a classification problem that classifies sentiment in two, three or four categories (positive, negative, natural and mixed). Since the problem resembles text classification, all the methods that apply there can be used. Those approaches depend on finding sentiment words that indicate positive or negative opinion. The simplest approach in this case it to use a bag of words representation as features; all other relationships between words are ignored. The drawback of this

method is that it does not take into consideration context cues and the fact that some words can be classified as negative in one situation and positive in another (El-Halees, 2011).

The first work that aims at detecting sentiment from text at the document level in Arabic used a supervised learning method (Abbasi, Chen, & Salem, 2008). In their work the goal was to build a sentiment analysis that worked for multiple languages. They experimented with Arabic and English. They used Support Vector Machines (SVM) and Entropy Weighted Genetic Algorithm (EWGA) to select features for both Arabic and English. The test was done on a small dataset from two extremist forums. One forum was in Arabic and the other was in English. They achieved 91% accuracy in the Arabic web forum and 90% on the English. They concluded that using both stylistic and syntactic features improved accuracy. Another supervised approach at the document level was used by Omar, Albared, Al-Shabi, and Al-Moslmi (2013). They used Arabic reviews collected from Jeeran.com – a popular Arabic review site for products and services, which is similar to Yelp in the United States. 3450 reviews were divided into a 3000 items training set and 450 items testing test. Items were divided into objective, positive, negative, neutrals reviews. Two annotators were used to annotate; college-educated native speakers were used to resolve any remaining conflicts. The approach used was a two-stage classifier. The first one classified the sentiment using three machine learning techniques: SVM, Naïve Bayes (NB) and Rocchio classifier as base classifiers. Ensemble voting method was then used as a fixed process and a meta-classifier to gather output through training methods. The evaluation shows that the NB algorithm out-performs other classifiers in subjectivity analysis and SVM (89.81) performs slightly better than Naïve Bayes (89.78) in sentiment analysis.

Another approach is to use combined classifiers instead of single classifiers. One such study uses a system which goes through three classifiers sequentially: the first is lexicon based

classifier that has a dictionary of positive and negative words manually constructed from translating SentiStrength to Arabic and using an online dictionary to add common Arabic words. Any document that left unclassified in this phase goes through a second phase which is based on a Maximum Entropy classifier that uses the previously classified documents as a training set. The remaining documents which are not classified in phase two are classified in the last phase based on K-nearest neighbor classifier (Thelwall, Buckley, Paltoglou, Cai, & Kappas, 2010). The approach was tested on a corpus that consists of three different domains: education, politics and sports. It contained 1143 posts with 9793 Arabic statements. Accuracy, precision and recall were used to evaluate the classifiers. The authors concluded that using three different classifiers performed better than using one or two classifiers and that positive statements performed better than the negative ones (f-score 81.7% for positive documents and 78.9% for negative documents over the three domains). The authors did not specify the source of the three genres used in the corpus and the academic grounds for choosing the sequence of classifiers was not explained (Thelwall et al., 2010).

Another line of work at the document level used classifier features specific to Arabic. Farra, Challita, Assi, and Hajj (2010) proposed a grammatical approach that makes use of Arabic sentence structure. They combined verbal and nominal sentences into a generic form. Subjects in both verbal and nominal sentences are considered actors and verbs are actions. They then created a training set that consisted of actors and actions labeled as features. The features used are sentence types, actor, action, object, adjective, type of pronoun and noun, transition, word polarity and sentence class.

Farra et al. (2010) also proposed a semantic orientation approach by extracting new features such as frequency of positive, negative and neutral words, the frequency of special

characters, the frequency of emphasis characters, the frequency of conclusive and contradiction words and other similar features. The system makes use of an interactive learning dictionary which stores the semantic polarity of word roots extracted by a stemmer and asks the users about any new word if it does not exist in the dictionary. In their evaluation, they used 29 sentences that were annotated manually with part-of-speech tags. They reported 89% accuracy with SVM classifier with 10-fold cross-validation. Also, they used 44 random documents for evaluating both of their approaches on J48 decision tree classifier. 62% accuracy was achieved with the semantic approach with the learning dictionary.

2.2.2.2 Sentence level Arabic Sentiment Analysis: Instead of identifying the overall sentiment of the document, sentence level analysis differentiates between subjective and objective sentences and then identifies the sentiment orientation of each subjective sentence. Subjectivity classification is beyond the scope of this proposal because it is not essential to aspect extraction. For a survey of subjectivity classification work in Arabic refer to (Korayem et al., 2012)

Abdul-Majeed and his colleagues have done a series of experiments on sentiment analysis at the sentence level that developed a corpus of 2855 sentences from the Penn Arabic Tree Bank (Abdul-Mageed & Diab, 2012a, 2014; Abdul-Mageed & Korayem, 2010; Abdul-Mageed, Kübler, & Diab, 2012). Those sentences were annotated by two college-educated native speakers of Arabic. Each sentence was identified as Objective and Subjective (Positive, Negative and Neutral). In their studies, they used many features: language independent features, Arabic-morphological features and genre-specific features. They also studied the level of stemming required for such a system and found that stem setting out-performs other lemmatization used. The system classifies subjectivity and sentiment in news wire data written in MSA at the sentence level. The system

used an SVM classifier for subjectivity followed by an SVM classifier for sentiment. They report 95.52% accuracy using unique, domain and adjective features (Abdul-Mageed & Diab, 2011)

In another study (Abdul-Mageed et al., 2012) used SVM based approach on SAMAR a system for subjectivity and sentiment analysis for Arabic (SSA). The aim of the research was to experiment with answering the following research questions: how to best represent lexical information? Are the standard features useful? How to deal with Arabic dialect and to what extent do genre-specific features impact performance? The features used are word form, POS tagging, UNIQUE tag, polarity lexicon, genre specific feature (gender, user ID). The evaluation that the highest accuracy reached for subjectivity classification was about 96% using token, polarity, POS tags as features. The highest for sentiment classification was about 71% using polarity, gender and lemma as features.

Work by Elhawary and Elfeky (2010) was also at the sentence level but with the aim of incorporating the work in a search engine. They worked with business reviews extracted from the web. The sentiment analysis classifier developed was based on previous work on English sentiment analysis (Blair-Goldensohn et al., 2008). Their analysis shows high precision for positive and negative sentiment but a lower precision for mixed and neutral sentiment which is the same case in English.

An unsupervised approach at the sentence level was used by Al-Subaihin, Al-Khalifa, and Al-Salman (2011). They used a gaming approach to analyze sentiment on a corpus extracted from an Arabic restaurant review site. The reviews were written in Arabic dialect. The system is based on two steps; the first one is a gaming approach to build the lexicon dictionary through player annotations. The second step is a sentiment analyzer which works through word segmentation and

then calculates the overall sentiment score for the review. The highest precision reached by their system is 60.32%.

2.2.2.3 Aspect Level Sentiment Analysis and its Subtasks: In order to do sentiment analysis at a deeper level than document and sentence level, there are some sub-tasks involved in the process. For English, there are many studies that analyze opinions at this level including extracting, categorizing, summarizing aspects and extracting opinion holder or opinion target. The lack of reliable publicly available Arabic NLP tools makes many of these tasks harder. There are a few preliminary works at this level on Arabic. There is one work for Arabic opinion holder extraction which was done by Elarnaoty et al (2012). There is a preliminary work on establishing the basis for Arabic Aspect based Sentiment analysis. The International Workshop on Semantic Evaluation 2016 (SemEval-2016) has established a data set for the task of aspect sentiment analysis with a sub-task for aspect extraction but the task for Arabic aspect based extraction has not received any submission(Pontiki et al., 2016). Al-Smadi, Qawasmeh, Talafha, and Quwaider (2015) have recently established aspect based annotation for a book review dataset (LABR). The data set has a baseline for aspect sentiment classification tasks but to the best of our knowledge is no published work on aspect extraction methods on similar Arabic reviews. In this work, we aim at applying all the methods that have been applied to English reviews for the aspect extraction task and evaluate their performance on translated Arabic reviews. In this research, we will also establish the basis for the work at the aspect level for Arabic sentiment analysis by carrying out some experiments

with supervised and unsupervised approach previously applied to English reviews. We will also develop a data set that can be used later for similar tasks.

2.2.2.4 Arabic Sentiment Analysis in Social Media The work in this area is limited to preliminary studies. Ahmed, Pasquier, and Qadah (2014) did a preliminary study to examine how preprocessing of Arabic social media text can improve sentiment analysis. The study was done on Twitter through crawling tweets on four keywords: “Obama”, “Messi”, “Iphone”, and “shia”. They recommend using unigrams which had the highest accuracy in their data. The preprocessing they recommend consists of tag adding and normalization. The tags that were added are URL, USERNAME, HASHTAG, SAD, HAPPY, LAUGH and TARGET. Through using SVM, Naïve Bayes, Maximum Entropy, Bayes Net and J48 decision tree they test the effect of the previous steps of preprocessing, they reported the impact on improving sentiment accuracy. The five machine learning classifiers used had comparable results.

Another study by Albraheem and Al-Khalifa (2012) examined the problems of SA in Dialect Arabic. The experiment was done on 100 Tweets extracted from Saudi hashtags in Twitter. The study did not specify what those hash tags were. They implemented a simple system to analyze sentiment by first tokenizing, preprocessing and stemming each tweet. Then, each token is looked up in a lexicon based dictionary to determine its polarity. A simple polarity sum was then used to determine the polarity of each tweet. The paper does not state the source of the lexicon based dictionary that was used to determine the polarity of each sentence.

Mourad and Darwish (2013) provided a new annotated dataset for Arabic tweets that consists of 2300 tweets based on expanding current SSA lexicon corpus (Arabic and English translated to Arabic). Their system combines a wide variety of features from previous works in

SVM and Naïve Bayes classifier and shows some improvement in using a large corpus compared to previous research.

Soliman, Elmasry, Hedar, and Doss (2014) collected 1846 comments from four news websites: Aljazera, BBC Arabic, Alyum Alsabe, and Al-Arabiya, as well as two Facebook pages. The work focuses on both MSA and DA, more specifically, the Egyptian Dialect Arabic. The approach used a lexicon based on the previous available corpora developed in (Abdul-Mageed & Diab, 2012b; Rushdi-Saleh, Martín-Valdivia, Ureña-López, & Perea-Ortega, 2011) and added 43 words and 27 idioms from their dataset. SVM based classifier was used for subjectivity analysis.

2.2.2.5 The Use of Machine Translation (MT) for Sentiment Analysis Cross-lingual information transfer is very common in NLP applications to overcome resource scarcity of one language with resources from a wealthy language. Arabic is a language that lacks the availability of NLP resources especially for processing DA. English is a language that has been studied extensively and has a lot of resources that have been applied to different language through the use of machine translation (MT). MT has been applied to translate Arabic to English for the task of opinion mining. Rushdi-Saleh et al. (2011) experimented with the effect of translating opinion mining movie corpus called Opinion Corpus for Arabic (OCA) from MSA to English and concluded that although a slight loss of precision due to translation existed (-0.43%), the results were comparable to the results using English text. Abdul-Mageed and Diab (2014) also used MT to leverage English resources to large scale multi-genre multi-dialectal multi-lingual lexicon for the purpose of subjectivity and sentiment analysis. Refaee and Rieser (2015) compared the performance of SA approaches with and without translation on Social media text which consist mostly of DA. They concluded that MT provides a cheap and effective way to build an SA system. In this research, we take advantage of translating our dataset to English and leverage the wealth of

NLP resources and also to compare the performance of translated reviews along with English NLP tools to the performance of Arabic reviews along with Arabic NLP tools

3.0 CHALLENGES

There are two types of challenges faced when conducting opinion mining in Arabic: some challenges are related to the field of opinion mining and others are related to the Arabic language.

3.1 CHALLENGES RELATED TO THE NATURE OF OPINIONS

Opinion mining faces several problems. First, identifying the set of words that can identify the polarity in the text is generally hard. Many adjectives are domain dependent (ex: The battery life is long vs takes a long time to boot). Similarly, sentiment and subjectivity are context sensitive. The sentence “reading the book was very enjoyable” is negative in the movie review context but positive in the book review context. Finally, some opinions are expressed in idioms and not individual words (e.g. cost an arm and leg)

The last problem is that semantics depends a lot on the word sequence and sentence structure. Saying “Mac is more expensive than Windows” is not the same as “Windows is more expensive than Mac”. Unfortunately, this problem has not been explored for Arabic language. Ahmed et al. (2014) relates the lack of reliable Arabic NLP resources such as reliable syntactic parser as the main problem for not exploring this problem.

3.2 CHALLENGES RELATED TO NATURE AND USAGE OF ARABIC LANGUAGE

The Arabic Language is divided into three types: Classical Arabic, Modern Standard Arabic (MSA) and Dialect Arabic (DA) (Soliman et al., 2014). The Arabic language has many different dialects that are used in informal daily communications but are not standardized or taught formally in schools. While there are a variety of dialects, MSA is the only one standard form that is widely recognized and formally taught in schools. MSA is based on Classical Arabic which is the language of the Qur'an (Muslims' holy book) (Habash, 2010). The MSA is not a native language of any country and it is largely different from dialect forms. MSA has been studied extensively and many NLP tools are available for it. Unfortunately, most of the web contents are written in the dialectal form that has not been studied as much. To the best of our knowledge there exist no reliable NLP tools for it.

Arabic is a morphologically rich language (MRL) where most of the information regarding syntax and relation is expressed at the word level. English on the other hand has much less information expressed at the word level. The Arabic base form of a word can lead to thousands of surface forms while in English a verb would have three different forms so using those forms in a lexicon corpus will lead to data sparseness in Arabic while in English there is a high chance that the three terms will be present in text (Abdul-Mageed et al., 2012; Ahmed et al., 2014). This suggests using a compact form of the word along with POS tagging to overcome the problem of data sparseness. Albraheem and Al-Khalifa (2012) also recommend stemming to reduce the size

of the lexicon corpus. On the other hand, (Rushdi-Saleh et al., 2011) does not recommend the use of stemming for the task of opinion mining.

The second challenge related to Arabic is the lack of widely available Arabic corpora (Abdul-Mageed & Diab, 2012a), the lack of Arabic lexicon that can be used for sentiment and the lack of publicly available and reliable NLP tools such as Part of speech tagger and dependency parser.

Opinion mining has gained a lot of popularity with the rise of social media. The amount of user-generated content available for researchers has increased tremendously but Arabic opinion mining has received little attention compared to English language. While the amount of data provides many opportunities for researchers, the data are highly unstructured and contains misspellings, abbreviation, repetitions “sooooo Happyyyyyy” and concatenated words (Albraheem & Al-Khalifa, 2012; Joshi, Balamurali, Bhattacharyya, & Mohanty, 2011).. Also, most of the Arabic content on the web is written in the dialect form.

The use of informal form on the web content leads to many problems. Arabic users encode Arabic words in roman alphabet for example “ الحرب which means “war” is written as “Al7arb ” or “Al 7arb” and there is no defined standards about how this is done so each word would have different variations depending on the user(Ahmed et al., 2014). Also, Albraheem and Al-Khalifa (2012)in their study of problems related to DA, indicated that different words with different meaning have the same root which can impact SA if the wrong root have different sentiment. Appendix A covers more detail about the nature of Arabic language.

4.0 DATASET

The successful completion of sentiment analysis tasks depends on the availability of a dataset for training and testing. In the case of aspect sentiment analysis in Arabic, there is very limited supply of datasets created for this task. There is large set of book reviews created for aspect based sentiment analysis by Al-Smadi et al. (2015) but the dataset was not publicly available. Furthermore, we believe that the SEMEVAL competition may lead to a widely accepted dataset. In 2016, SEMEVAL competition created a task specific for aspect based sentiment analysis that included an Arabic hotel reviews dataset. This dataset was extracted from Booking.com. Unfortunately, the task received no submission for Arabic aspect extraction. The dataset is publicly available (Pontiki et al., 2016). We used their training set to compare the performance of the aspect extraction methods to the performance on our dataset.

Although there is some effort in establishing the aspect level sentiment analysis, to the best of our knowledge, there is no restaurant reviews dataset that is labeled specifically for this task. Consequently, we created a restaurant review dataset that can be used for aspect sentiment analysis and aspect extraction to complement SEMEVAL hotel review dataset.

Our dataset was collected from Jeeran.com which is one of the popular Arabic sites for services' reviews. The data are structured. Spams are eliminated from the reviews by the website owner. Categories varies widely some of which include shopping, restaurants, travel, financial services, etc. The next section describes the collection, preparation and description of the dataset.

We will refer to this dataset as restaurant reviews dataset (RR) and we will refer to the translated version of this dataset as (TRR).

4.1 THE RESTURANT REVIEW DATASET (RR)

4.1.1 Data Collection

The data used in this research were crawled between May and July 2014. The website covers cities in Saudi Arabia, Jordan, United Arab Emirates, Egypt, Kuwait and Qatar. We limited the crawls to cities in the Kingdom of Saudi Arabia to restrict the wide variety of dialects available in the web site. Most of our dataset contains Gulf Dialect Arabic with some MSA and English reviews.

Our goal for crawling was to collect a sample of reviews and not necessarily all the reviews. Reviews were collected on a per-place basis and limited to the Riyadh city. The total reviews crawled 6485 for the restaurant domain. 500 random reviews were selected to develop the corpus. The following sections (4.1.2-4.1.3) describe the steps we took to develop the corpus used for the task of aspect extraction.

4.1.2 Preprocessing

Text on the web generally is unstructured and, particularly in the case of Arabic, is a bit messy. To facilitate the task, we excluded reviews written in English because the focus of this research is on the Arabic dialect and not English reviews. A common problem with DA is that users tend to write the same word in different ways as shown in Figure 1. These spelling variations

cannot be easily fixed due to the lack of gold standard Arabic DA dictionary. To alleviate the problem, we resorted to simple preprocessing steps as follow:

1. Remove punctuation, diacritics, and any non-characters.
2. Normalize the Arabic letter أ إ آ with ا
3. Remove the extension from the words for example بـــــــاب is reduced to باب
4. Replace ى with ي
5. Replace ة with ه
6. If a word starts with ء then replace it with ا
7. Replace ؤ with و
8. Replace ىء and ئ with ي

These steps have been used in the literature before for similar tasks and it has been shown to increase accuracy for sentiment analysis (Ahmed et al., 2014). Although these steps will not overcome the spelling variations, it will help alleviate the problem by reducing the form for words that have some similar variations as in Figure 1(a, b, c, d) which have spelling variation in one letter or diacritics. The problem of Figure 1 (e, f) are for a sequence of different letters and thus cannot be solved easily without access to gold standard dictionary.

(a)	Eat	أكل	اكل	
(b)	Meal	وجبه	وجبة	
(c)	Very	جداً	جدا	
(d)	Hungry	جوعانه	جيعانه	جعانه
(e)	Bad	سيئ	سيء	
(f)	Light	ضوء	ضوا	ضو

Figure 1. *Arabic dialect variety of words spelling*

4.1.3 Labeling

After the preprocessing, we need labels that will serve as a training and testing dataset. The labels are considered the gold standard to the system. For the task of Aspect extraction using the supervised method CRF, we have created labels that represent the three main parts of speech of the Arabic vocabulary. Those labels are explained in section 4.1.3.1. For the task of Aspect extraction and sentiment classification for both the supervised and unsupervised method we have created Aspect and sentiment labels as explained in section 4.1.3.2 and 4.1.3.2

4.1.3.1 Part of Speech Labeling: The task of Aspect extraction using CRF requires the use of labels that serve as features to the model. There are a variety of features that can be used. Our selected features and how they are implemented are explained in details in section 5.4.2.1. Since Arabic part of speech consist of three main part of speech (Nominal, Verb and Particle), we asked

the annotators to label the data with those three main part of speech and we refer to them as Super POS tags (SPOS) to distinguish them from the POS derived from Arabic POS tagger. Those three part of speech are:

- **Nominal (N):** Nouns (Noun, Proper Noun), Derived nouns (Adjectives, Imperative verbal noun), Personal Pronoun, Demonstrative pronouns اسماء اشارة, Possessive determiners الضمائر المتصلة, Relative pronouns الاسماء الموصولة, Adverbs (Time adverbs, location adverbs)
- **Verb (V)**
- **Particle (P):** Prepositions such as: from, to, in; من, الى ، في Subjunctive particles such as: لن ، -ل أن، Jussive particles such as: لم ، لا, Negative particles such as: لا ، ما
- **Adjectives (Adj):** subcategory of Nominal

We have also added a fourth label which is **Others (O)** that is used for any other sequence of characters (not words) that does not fit the other categories. Table 1 provides statistics of these labels. The inter-annotate agreement (Kappa) is 0.877. The source of the disagreement between the annotator is due to misspelling and lack of structure in some of the sentences.

Table 1. SPOS Label Statistics

	N	**V**	**P**	**Adj**	**O**
Total words	5888	1147	1582	1435	215
Distinct words	2617	733	202	575	55

4.1.3.2 Aspect-Sentiment Labeling: As the gold standard for the training and testing both supervised and unsupervised, we randomly selected 500 reviews which were then annotated by two graduate students who are native Arabic speakers. Labeling the corpus was done in three steps.

First, for the task of aspect extraction, each annotator viewed a sentence from the restaurant domain and marked each explicit aspect discussed in the review. We used the Inside-Out-Beginning IOB labeling scheme, if the aspect consists of one word, the label B-Asp is used. If the aspect consists of more than one word, the label B-Asp is used for the first word and I-Asp is used for the subsequent words. All non-aspect terms are labeled with O. If the sentence has no explicit aspects, the annotators were asked not to add any labels.

Second, for the task of sentiment classification, we show the annotators a sentence with the aspect word underlined, the annotators were asked to identify the words that express sentiments directly to the aspect underlined and apply a 'SENTMENT' tag to it. They were also asked to specify each sentiment label as positive (POS), negative (NEG) or neutral (NEU). The inter-annotate agreement (Kappa=0.9). The source of disagreement between annotators is due to misspelling or confusion between parts of speeches. 500 reviews are labeled with 2261 aspect (1000 distinct aspects). A summary of data statistics is provided in Table 2, Table 3 and Table 4.

Table 2. Aspect Statistics

Reviews	# of tokens Per review	Explicit Aspects	Average Aspect Per Review
500	10313	2263	2.9

Table 3. Aspect-Sentiment Statistics

Tagged reviews	Tokens Per review	Explicit Aspects	Explicit Aspects With Sentiment	Explicit Aspects without sentiment
500	10313	2263 (avg 2.9)	1186	1163

Table 4. Aspect- Sentiment distribution

	Positive	Neutral	Negative
Aspects with sentiment	909	55	234

Figure 2 shows an example of review labeled for the task of aspect extraction. We followed XML format similar to SemEval dataset to facilitate comparison between the three datasets. Note that RR is the only dataset that contain SOPS label as explained in section 4.1.3.1

```
<Review rid="456">
 <text>الحمص لذيذ والعصائر الطازجه لذيذه</text>
<SPOS>
<word position="0" label="N"/>
<word position="1" label="adj"/>
<word position="2" label="N"/>
<word position="3" label="adj"/>
<word position="4" label="adj"/>
</SPOS>
<Opinions>
<Opinion aspect="الحمص" polarity="positive" position="0" opinionword="لذيذ" owposition="1" type="B-asp"/>
<Opinion aspect="والعصائر" polarity="positive" position="2" opinionword="لذيذه" owposition="4" type="B-asp"/>
<Opinion aspect="الطازجه" polarity="positive" position="3" opinionword="لذيذه" owposition="4" type="I-asp"/>
</Opinions>
</Review>
```

Figure 2. Example of RR labeled review

4.2 TRANSLATED RESTAURANT REVIEW DATASET (TRR)

Translation is a common method used in NLP task for low resource languages such as Arabic. For our purpose, we used translation as a method to utilize the rich NLP resources available for English and use it for low resource languages (Arabic in our case). We employed Microsoft Bing translator to create a translated version of our restaurant reviews dataset (RR). The Microsoft Bing translator was used because it outperformed Google Translator for sentiment analysis for DA in previous work (Refaee & Rieser, 2015). We preserved aspect and sentiment labels during the translation process which facilitated the use of this dataset for aspect extraction. Figure 3 shows an example of a translated review from TRR. We used a similar XML format to the one used in SemEval dataset to facilities comparisons between the datasets.

```
<Review rid="456">
        <text>الحمص لذيذ والعصائر الطازجه لذيذه</text>
        <translation> Chickpea delicious and juices fresh delicious </translation>
<Opinions>
<Opinion aspect="Chickpea" polarity="positive" position="0" opinionword="delicious" owposition="1" type="B-asp"/>
<Opinion aspect="and" polarity="positive" position="2" opinionword="delicious" owposition="5" type="B-asp"/>
< Opinion aspect="juices" polarity="positive" position="3" opinionword="delicious" owposition="5" type="I-asp"/>
<Opinion aspect="fresh" polarity="positive" position="4" opinionword="delicious" owposition="5" type="I-asp"/>
</Opinions>
</Review>
```

Figure 3. Example of TRR labeled review

4.3 2016 INTERNATIONAL SEMANTIC EVALUATION WORKSHOP DATASET (SEMEVAL)

SemEval 2016 workshop introduced the task of aspect extraction for sentiment analysis for Arabic dialect. Unfortunately, the task received no submission for the Arabic. Regardless, they contributed a large Arabic hotel reviews dataset labeled for aspect extraction, aspect sentiment identification and also aspect categorization (Pontiki et al., 2016). We used SemEval Arabic hotel review training dataset for our task (Arabic_Hotels_TrD_V2.xml). The dataset consists of 1839 reviews which are label for the mentioned tasks. Figure 4 shows an example of review number 456 obtained from the SemEval Dataset. The dataset labeled for aspect extraction, aspect categorization and aspect sentiment analysis. It For more information about the dataset refer to Pontiki et al. (2016). Section 4.4 provide dataset statistics and compare it to RR and TRR.

```
<Review rid="456">
 <sentences>
 <sentence id="456:0">
 <text>أنصح بالنوم وليس تناول الطعام  موقع مثالي للإقامة قبل رحلة طيران مبكرة.</text>
<Opinions>
<Opinion target="موقع" category="LOCATION#GENERAL" polarity="positive" from="31" to="35"/>
</Opinions>
</sentence>
<sentence id="456:1">
<text>كانت الغرفة ممتازة وكذلك الموظفون وبوفيه الإفطار. ومع ذلك فقد كانت وجبة العشاء في المطعم باهظة الثمن وغير مرضية.</text>
<Opinions>
<Opinion target="الغرفة" category="ROOMS#GENERAL" polarity="positive" from="5" to="11"/>
<Opinion target="الموظفون" category="SERVICE#GENERAL" polarity="positive" from="25" to="33"/>
<Opinion target="بوفيه الإفطار" category="FOOD_DRINKS#QUALITY" polarity="positive" from="35" to="48"/>
<Opinion target="وجبة العشاء" category="FOOD_DRINKS#PRICES" polarity="negative" from="67" to="78"/>
</Opinions>
</sentence>
<sentence id="456:2">
<text>عند الوصول في المطار بدلا من ذلك M _ S فندق يتميز بمرافق نوعيّة وخلّاقة وساخنة. قم بشراء وجبة داخل الغرفة من</text>
<Opinions>
<Opinion target="فندق" category="HOTEL#QUALITY" polarity="positive" from="0" to="4"/>
</Opinions>
</sentence>
</sentences>
</Review>
```

Figure 4. Example of SemEval labeled review

4.4 SUMMARY OF DATASETS

We have gathered three datasets to be used in the task of aspect extraction. These datasets are Arabic Restaurant Reviews (RR), Translated Restaurant Review (TRR) and SemEval 2016 (SemEval) datasets. RR is a restaurant reviews dataset developed in house specifically for the task of aspect extraction. RR was translated to TRR using Bing Microsoft translator. The goal of the translation is to compare the performance of aspect extraction methods using translation to using Arabic NLP tools. SemEval is also used to compare the performance of the methods on two Arabic dialect datasets. Both SemEval and RR contains Arabic dialect review but in developing RR we limited the crawling to the city of Riyadh to limit the variety of dialects in the website. Thus, RR contains mostly Gulf Dialect. In comparison, SemEval was collected from reviews for hotels in different cities around the Middle East, examining the dataset we noticed it contains a mix of dialects as well as Arabic MSA. We also observed that SemEval reviews are longer than RR reviews (54.74 tokens per review compared to 25). The length of the reviews led to slightly more aspect per review than RR (5 compared to 4.53).

In terms of labeling, we adopted IOB labeling scheme as shown in Figure 2 and Figure 3 while SemEval dataset used a different scheme as shown in Figure 4. There is no effect on the labeling method in the result of this research. Although RR and SemEval were both created for Aspect sentiment analysis tasks, SemEval labeling extend the task to cover aspect categorization as well. While both SemEval and RR support aspect sentiment analysis by identifying aspect sentiment, RR goes a deeper level by identifying direct opinion words related to the aspect as shown in Figure 2. Example of RR labeled review. This direct relation between sentiment word

and their aspect was also preserved in TRR. Table 5 and Table 6 provide a comparison of statistics of the three datasets.

Table 5. Datasets statistics

Dataset	Total reviews	#sentences	Total tokens	Tokens per review	Tokens per sentence
RR	500	-	10313	25	-
TRR	500	-	11390	22.78	-
SemEval	1839	4802	100683	54.74	20.96

Table 6. Datasets aspect statistics

	Reviews	Explicit aspects	Distinct aspects	Aspect/review
RR	500	2263	1000	4.53
TRR	500	2548	880	5.0
SemEval	1839	9760	1616	5.3

5.0 RESEARCH DESIGN

This research evaluates opinion mining for the Arabic social media text at a fine-grained aspect level. The limited resources available to process Arabic dialectal text make the task harder. We utilized the resources available to richer languages such as English through the use of machine translation. We built two datasets that can be used for such tasks. We then applied the methods that have been employed for English opinion mining at the aspect level and that can be applied to Arabic reviews without developing extensive NLP resources. Those methods are one supervised methods explained in section 5.4, one semi-supervised method explained in section 0, and two unsupervised methods explained in sections 5.6 and 5.7.

5.1 ASSUMPTIONS

In the case of general text, one would need to do subjectivity analysis before proceeding with the next step of doing sentiment analysis. In this study, we assume that all the sentences contained in the review are subjective – reviews usually contain the opinion of the users on the item or service being reviewed.

5.2 DELIMITATION

This study examines the form of Arabic that is widely used on the web and social media that correspond to the spoken form of Arabic (Arabic dialect). This study examines the methods that can be used to extract explicit aspects from the domain of restaurant reviews which has been used in other languages. Some of these methods cannot be explored in the context of Arabic because of the lack of reliable NLP tools such as morphological analyzer and the lack of reliable dictionary. This study is limited to the restaurant domain. Topic models are unsupervised, thus do not exhibit domain dependency. We limited the experiment to 500 of reviews because of the effort required for labeling sentences. This study is limited to identifying explicit aspects and thus we are not looking into implicit aspects.

5.3 LIMITATIONS

Conditional random field and Topic modeling are domain specific and may or may not perform at the same level on a different domain. The lack of reliable NLP tools for Arabic dialect and the challenge that exist for the use of social media text of Arabic limit the performance of the system. We exhibit a loss of precision due to the use of machine translation between Arabic dialect and English.

5.4 ASPECT EXTRACTION USING CONDITIONAL RANDOM FIELD (CRF)

5.4.1 Background

CRF uses a discriminative undirected probabilistic graphical model. They were first introduced by Lafferty et al. (2001) for the task of labeling sequential data for speech recognition tasks. It is used to model known relationships between observations and then construct consistent interpretation. It is widely used in sequence labeling problems, i.e. Natural Language Processing such as part of speech tagging, Named Entity Recognition (NER), and Information Extraction (IE). It was also used in other problems such as biological sequencing, image and video labeling, and image recognition. Similarly, Linear-chain CRF has been applied to Aspect extraction tasks because the problem of finding aspects in a sentence can be viewed as a sequence labeling problem.

CRF is a generalized form of Hidden Markov Model. Formally, given a sequence of tokens (observations) $x = x_1 x_2 x_3 .. x_n$, we need to generate a sequence of labels (hidden states) $y = y_1 y_2 y_3 .. y_n$ for each token x. For our purpose, the set of possible labels are ASPECT and NON-ASPECT. The aim of CRF model is to find y that maximizes $p(x|y)$ for the given sequence.

$$p(y|x) = \frac{1}{z(x)} * \exp(\sum_t \sum_k \lambda_k f_k(y_{t-1}, y_t, x))$$

$$z(x) = \sum_{y \epsilon Y} \exp(\sum_t \sum_k \lambda_k f_k(y_{t-1}, y_t, x))$$

$$\lambda_k \; is \; the \; weight \; of \; f_k$$

While CRF and Hidden Markov Models (HMM) are sequence modeling techniques, CRF overcomes the limitation of by relaxing the independent conditional assumption which is given the hidden state, observations are independent. Thus, HMM ca not model the interaction between adjacent tokens which CRF does.

5.4.2 Applying CRF to Arabic Dialect

We applied 10-fold cross validation CRF to RR and TRR for the task of aspect extraction as a supervised method. CRF relies on a defined set of features that are fed to the model during training. We created two sets of those features one for RR and the second is for the translated version of those reviews (TRR). MSA Arabic NLP tools were used to create the set of features to be used for Arabic CRF models. More precisely, we employed MADAMIRA (Pasha et al., 2014) as a POS tagger and NER tool. We also used CAMEL (Habash & Roth, 2009; Marton, Habash, & Rambow, 2010, 2013) as an Arabic dependency parser. We then used Stanford Core NLP (Manning et al., 2014) to derive the set of features to be used in our CRF model. Those set of features are defined in the following section.

5.4.2.1 CRF Features: CRF models depend on a set of features that are fed to the model. We created two similar sets of features: one for the Arabic Restaurant review dataset and one for the translated restaurant reviews. Those features are described below:

- **Part of Speech tags (POS):** a set of POS tags applied through the use of part of speech tagger. We employed Stanford POS tagger from the Stanford NLP Core suit to extract POS tags for TRR. In a similar fashion, we employed MADAMIRA as a POS tagger for RR.

- **Super Part of Speech (SPOS):** Although the use of the upper classes of Arabic Part of Speech have not been experimented with in the literature, we would like to make use of it to support the POS produce by MADAMIRA and we refer to them as (SPOS). These labels are explained in section 4.1.3.1. These tags were only used for Arabic review because there are no equivalent tags for English reviews and during the translation used is not word to word translation where these labels can be reserved.
- **Named Entity (NE):** Entities in the review identified by Stanford Core NLP Named entity tagger for TRR. Similarly, they were identified by MADAMIRA for Arabic review dataset (RR). We refer to the translated name entities as NE.
- **Short dependency path (SP):** Previous works have successfully employed the use of dependency parser to extract direct relations to opinion expression (Jakob & Gurevych, 2010; Zhuang et al., 2006). More specifically “amod” and “nsubj” were extracted because they have been shown that they are the most accurate relationship between opinion and aspects. We used Stanford dependency parser to implement this feature for TRR dataset and we refer to them as SP. On the other hand, we used CAMEL dependency parser to implement this feature for RR dataset.
- **Word distance (WD):** the closest noun phrase to sentiment word (Jakob & Gurevych, 2010). We used Stanford POS tagger to find this feature for TRR and we refer to it as WD. On a similar way, we used MADAMIRA tagger for RR.
- **Sentiment words (SW):** the sentiment words are identified by our annotators and they are the words that holds opinion regarding the aspect being reviewed (section 4.1.3.2). We refer to this feature as SW.

  Figure 5 shows an example of how CRF labeling work on translated sentence from TRR. The same labeling method works on RR as well.

	Reading direction ←								
Review	الستيك	عجبني	شيء	أكثر	و	هادئ		المطعم	
Translated review	Steak	Like		Most	And	Quite	Is	restaurant	the
POS	NN	IN		JJS	CC	PDT	VBZ	NN	DT
NE	NO	NO		NO	NO	NO	NO	NO	NO
SP	YES	NO		NO	NO	NO	NO	YES	NO
WD	YES	NO		NO	NO	NO	NO	YES	YES
SW	NO	YES		NO	NO	YES	NO	NO	NO
Output	**B**	**O**		**O**	**O**	**O**	**O**	**B**	**O**

Figure 5. Example of CRF features on TRR

5.5 DOUBLE PROPAGATION ASPECT EXTRACTION

5.5.1 Background

Double propagation is a semi-supervised aspect extraction method developed by Qiu, Liu, Bu, and Chen (2011). The method is based on the idea that each opinion is related to a target in the sentence due to the fact that opinions are expressed on targets. The method assumes that targets are noun and noun phrases and opinion words are adjectives. It is considered semi-supervised because of the use of initial opinion seed words. This approach relies on identifying the syntactic relation on a sentence using a sentence parser which leads to the extraction of opinion words and targets by identifying certain relations.

This approach employs dependency grammar to define the syntactic relationship between words in a sentence. There are two type of dependency relations between words in a sentence.

Direct dependency where two words A and B depend directly on each other or A and B both depend directly on word H as illustrated in Figure 6(a) and (b). Indirect dependency where A depend on B through some additional words or A and B both depend on C through additional words in between as illustrated in Figure 6(c) and (d).

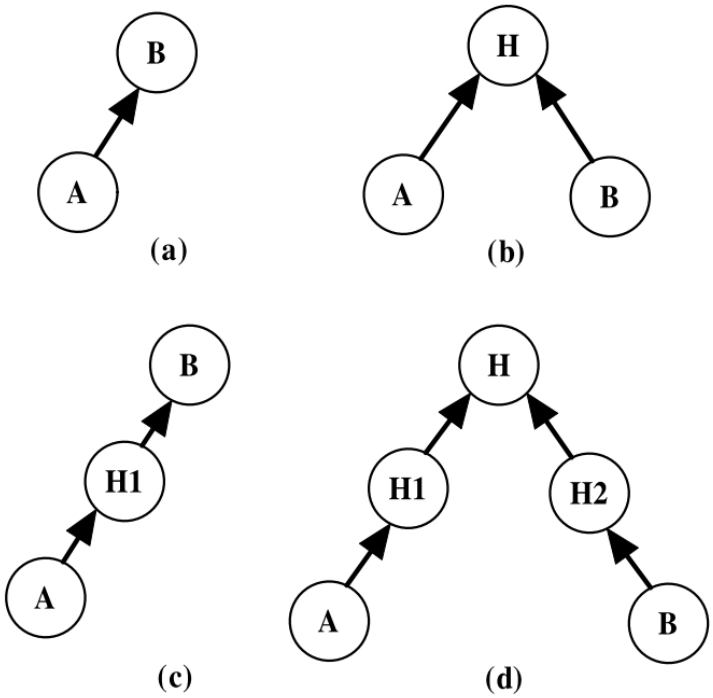

Figure 6. Different dependencies between words A B (Qiu et al., 2011)

Those dependencies described above are too general. They are then restricted by the authors to certain relations through the use of POS tagger and sentence parser. Aspects are restricted to noun and noun phrases identified by POS tags. Similarly, opinions are restricted to adjectives only. The only relations considered between aspect and opinions from the parser are *mod, pnmod, subj, s, obj, obj2 and desc*. The only relations considered between aspect words and

opinions themselves are *conj*. After identifying dependencies, POS tags and relations. The double propagation looks for certain relations that the actual extraction relies on. Those relations are shown on

Table 7.

Table 7. Rules for target and opinion word extraction (Qiu et al., 2011)

RuleID	Observations	output	Examples
$R1_1$	$O \rightarrow$ *O-Dep* $\rightarrow T$ *s.t.* $O \in \{O\}$, *O-Dep* $\in \{MR\}$, $POS(T) \in \{NN\}$	$t = T$	The phone has a good "screen". (*good* $\rightarrow$ *mod* $\rightarrow$ *screen*)
$R1_2$	$O \rightarrow$ *O-Dep* $\rightarrow H \leftarrow$ *T-Dep* $\leftarrow T$ *s.t.* $O \in \{O\}$, *O/T-Dep* $\in \{MR\}$, $POS(T) \in \{NN\}$	$t = T$	"iPod" is the best mp3 player. (*best* $\rightarrow$ *mod* $\rightarrow$ *player* $\leftarrow$ *subj* $\leftarrow$ *iPod*)
$R2_1$	$O \rightarrow$ *O-Dep* $\rightarrow T$ *s.t.* $T \in \{T\}$, *O-Dep* $\in \{MR\}$, $POS(O) \in \{JJ\}$	$o = O$	same as $R1_1$ with screen as the known word and good as the extracted word
$R2_2$	$O \rightarrow$ *O-Dep* $\rightarrow H \leftarrow$ *T-Dep* $\leftarrow T$ *s.t.* $T \in \{T\}$, *O/T-Dep* $\in \{MR\}$, $POS(O) \in \{JJ\}$	$o = O$	same as $R1_2$ with iPod as the known word and best as the extract word
$R3_1$	$T_{i(j)} \rightarrow T_{i(j)}$-*Dep* $\rightarrow T_{j(i)}$ *s.t.* $T_{j(i)} \in \{T\}$, $T_{i(j)}$-*Dep* $\in \{CONJ\}$, $POS(T_{i(j)}) \in \{NN\}$	$t = T_{i(j)}$	Does the player play dvd with audio and "video"? (*video* $\rightarrow$ *conj* $\rightarrow$ *audio*)
$R3_2$	$T_i \rightarrow T_i$-*Dep* $\rightarrow H \leftarrow T_j$-*Dep* $\leftarrow T_j$ *s.t.* $T_i \in \{T\}$, T_i-*Dep* $== T_j$-*Dep*, $POS(T_j) \in \{NN\}$	$t = T_j$	Canon "G3" has a great lens. (*lens* $\rightarrow$ *obj* $\rightarrow$ *has* $\leftarrow$ *subj* $\leftarrow$ *G3*)
$R4_1$	$O_{i(j)} \rightarrow O_{i(j)}$-*Dep* $\rightarrow O_{j(i)}$ *s.t.* $O_{j(i)} \in \{O\}$, $O_{i(j)}$-*Dep* $\in \{CONJ\}$, $POS(O_{i(j)}) \in \{JJ\}$	$o = O_{i(j)}$	The camera is amazing and "easy" to use. (*easy* $\rightarrow$ *conj* $\rightarrow$ *amazing*)
$R4_2$	$O_i \rightarrow O_i$-*Dep* $\rightarrow H \leftarrow O_j$-*Dep* $\leftarrow O_j$ *s.t.* $O_i \in \{O\}$, O_i-*Dep* $== O_j$-*Dep*, $POS(O_j) \in \{JJ\}$	$o = O_j$	If you want to buy a sexy, "cool", accessory-available mp3 player, you can choose iPod. (*sexy* $\rightarrow$ *mod* $\rightarrow$ *player* $\leftarrow$ *mod* $\leftarrow$ *cool*)

The double propagation method works as a bootstrapping method. It works as follow

1. Using the initial opinion seed words to extract aspect words using rules R1.
2. Use the extracted aspects to extract new aspects using rules R2.

3. Extract new opinion words through aspect extracted in step (1) and (2) using rules R3
4. Extract opinion words using both initial seed and extracted opinion words using rules R4

The detailed algorithm is described in details in Figure 7.

Input: Opinion Word Dictionary {O}, Review Data R
Output: All Possible Features {F}, The Expanded Opinion Lexicon {O-Expanded}
Function:
1. {O-Expanded} = {O}
2. $\{F_i\} = \emptyset, \{O_i\} = \emptyset$
3. for each parsed sentence in R
4. if(Extracted features not in {F})
5. Extract features $\{F_i\}$ using $R1_1$ and $R1_2$ based on opinion words in {O-Expanded}
6. endif
7. if(Extracted opinion words not in {O-Expanded})
8 Extract new opinion words $\{O_i\}$ using $R4_1$ and $R4_2$ based on opinion words in {O-Expanded}
9. endif
10. endfor
11. Set $\{F\} = \{F\} + \{F_i\}$, $\{O\text{-Expanded}\} = \{O\text{-Expanded}\} + \{O_i\}$
12. for each parsed sentence in R
13. if(Extracted features not in {F})
14. Extract features {F'} using $R3_1$ and $R3_2$ based on features in $\{F_i\}$
15. endif
16. if(Extracted opinion words not in {O-Expanded})
17. Extract opinion words {O'} using $R2_1$ and $R2_2$ based on features in $\{F_i\}$
18. endif
19. end for
20. Set $\{F_i\} = \{F_i\} + \{F'\}$, $\{O_i\} = \{O_i\} + \{O'\}$
21. Set $\{F\} = \{F\} + \{F'\}$, $\{O\text{-Expanded}\} = \{O\text{-Expanded}\} + \{O'\}$
22. Repeat 2 till size($\{F_i\}$) = 0, size($\{O_i\}$) = 0

Figure 7 double propogation algorithim. Reprinted from (Qiu et al., 2011)

5.5.2 Double Propagation Applied to Arabic Dialect

We followed the same method applied by Qiu et al. (2011). We applied this method to Arabic reviews and to their translation. We started by translating the reviews to English through Microsoft's translator Bing to facilitate the use of Stanford NLP tools. Then, Stanford POS tagger is applied to identify nouns, noun phrases and adjectives. Then, we applied MiniPar (D. Lin, 2003)

to extract those relationships (*mod, pnmod, subj, s, obj, obj2 and desc).* Then the same algorithm described in Figure 7 is applied. Stanford part of speech tagger was used for identifying N, NP and adj. We applied the same method in a similar fashion to Arabic reviews. We used CAMEL for the dependency parsing and MADAMIRA for the POS tagging. Note that both CAMEL and MiniPar use coarse set of dependency tags compared to Stanford comprehensive set.

We used the same set of seed words used by the original work that was adapted from Hu and Liu (2004a) for translated reviews (TRR). We adopted Arabic translation of the same set from the work by Salameh, Mohammad, and Kiritchenko (2015) to be used with the RR dataset.

5.6 ASPECT EXTRACTION USING FREQUENT NOUN AND NOUN PHRASES

5.6.1 Background

This method was proposed by Hu and Liu (2004a) and it is based on the notion that most users who reviews the same product use vocabulary that converges. Those vocabularies are the most frequent noun and noun phrases. This method tries to achieve aspect extraction in unsupervised fashion and by incorporating simple knowledge about sentence structure. The method can be divided in two parts. The first part is finding frequent based on the notion that aspects are the most discussed in the reviews and using an algorithm to find the most frequent aspects. The second part is based on a similar idea to double propagation that each aspect is coupled with opinion. hence, identifying all opinion words from the frequent aspects in the first step will lead us to find infrequent aspects by identifying the closest noun to those opinion words.

More precisely, the first step in this method is to identify the most frequent aspects that most customers identify. The way the authors determined those features is by doing POS tagging to find all noun and noun phrases since aspect are nouns that describe things related to the entity being discussed. A simple preprocessing if performed that includes removing stop words, stemming, and fuzzy matching for misspelling and word variations. Then, those noun and noun phrases are stored in a transaction file. Association mining based on Apriori algorithm are then performed to identify the candidate most frequent features. Finally, two types of pruning are performed to remove unlikely candidate aspects.

The first pruning step is compactness pruning dealing with aspect phrases which consist of more than one word. Because the input to the association mining algorithm has no indication of the word position on the actual sentence, the compactness pruning considers the position of the words in the sentence and the frequent aspect phrase should satisfy two conditions to remain a candidate.

- Condition 1: The frequent aspect phrase is compact when the word distance in the sentence is not greater than 3.
- Condition 2: If this phrase occur in more than one sentence in the reviews dataset, it should be compact (condition 1) on at least 2 sentences.

The second pruning step is redundancy pruning and deals with single word aspects by calculating the pure support (p-support) for aspect A. The p-support is the number of sentences that contain the aspect A as noun or noun phrase and the sentences that do not contain another aspect that is a superset of A. If the p-support of a candidate aspect A is less than 3 (based on the dataset that the author has), it is removed.

The remaining list of aspects after the pruning is the list of frequent aspects in the dataset. The next two steps are somewhat similar to double propagation and deal with extracting infrequent aspects in a very simple level. First, identify all opinion words based on the observation that "the closest adjective to noun or noun phrase is most likely opinion". Therefore, the closest adjective to the extracted frequent aspect is most likely opinion. Second, find the noun/noun phrase closest to all opinion word identified in the previous step, those noun and noun phrases are considered infrequent aspects. The authors argue that this simple processing produces good results because the infrequent features accounted for small percentage of all aspects 15%-20% (Hu & Liu, 2004b).

5.6.2 Aspect Extraction Using Frequent Noun and Noun Phrases Applied to Arabic Dialect

We used a modified approach based on the work of Hu and Liu (2004a) explained in the previous section. A similar process was used to extract aspect. We translated the Arabic dialect reviews to English using Bing translator. Then, we performed POS tagging using Stanford NLP system to identify all Noun and Noun phrases (Manning et al., 2014). Note that fuzzy matching that was used in the original work to correct spelling and deal with word variations were not performed here because we assume that the output of the translator is spelled correctly. Then, the Apriori association mining algorithm implemented by Borgelt (2012) was applied. We followed a similar pruning approach except that we determined the p-support threshold to be 2 based on our dataset for RR and TRR. Then, using the POS tags we identified all adjective that are closest to the identified aspects to be opinion words. Finally, all N and NP for those adjectives were considered infrequent aspects. This process is demonstrated in Figure 8. In a similar way, we used MADAMIRA to extract POS for Arabic reviews and we followed the same approach.

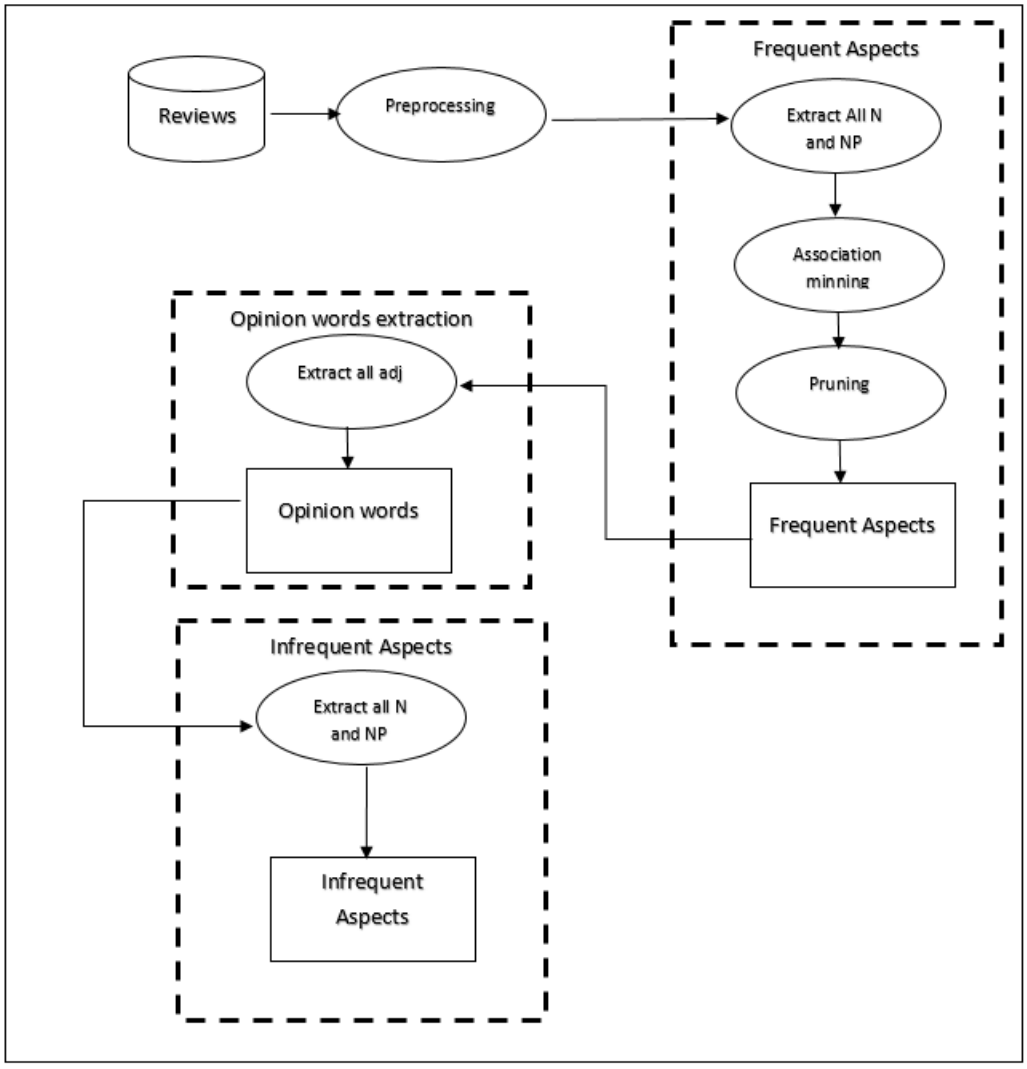

Figure 8. Frequent Noun and Noun phrases applied to Arabic

5.7 ASPECT EXTRACTION USING TOPIC MODELING

5.7.1 Background

Topic modeling is an unsupervised text mining approach to discover clusters of words called topics from a large collection of text documents. Each topic represents a probability distribution over words in the collection. Most of the aspect topic models are based on Latent Dirichlet Allocation (LDA). LDA takes set of Documents D as input. In our case, each review is considered a document. The model outputs are

- Document-topic distribution(θ): probability distribution over topics for each document.
- Topic-word distribution (ϕ): probability distribution over word for each topic

LDA assume that both $\theta \; and \; \phi$ follow multinomial distribution. Dirichlet prior are used to smooth the model. α is a Dirichlet prior on the per-document topic distributions and β is the parameter of the Dirichlet prior on the per-topic word distribution. Equal values for α and β are commonly used. This is called a symmetric Dirichlet distribution.

LDA works by the following algorithm:

- Let T {1,…,T be the number of Topics to be generated.
- V {1,…,V} is the number of unique words in the corpus.
- D is the number of documents
- Each documents d is a sequence of d_N
    - For each topic $t \in \{1, \dots, T\}$ do
    - Draw a word distribution for topic t, $\phi_t \sim Dirichlet\ (\beta)$
    - For each document $d \in \{1, \dots, D\}$ do
    - Draw a topic distribution for document d, $\theta_d \sim Dirichlet\ (\alpha)$
        - For each term $w_i, i \in \{1, \dots N_d\}$ Do
        - Draw a topic for the word, $z_i \sim Multinomial\ (\theta_d)$
        - Draw a word, $w_i \sim Multinomial\ (\varphi_{z_i})$

The graphical or plate notation for LDA model is given in Figure 9. The distribution used for θ *and* ϕ in our research is Gibbs sampling. For a more detail description of LDA refer to the original paper (Blei et al., 2003).

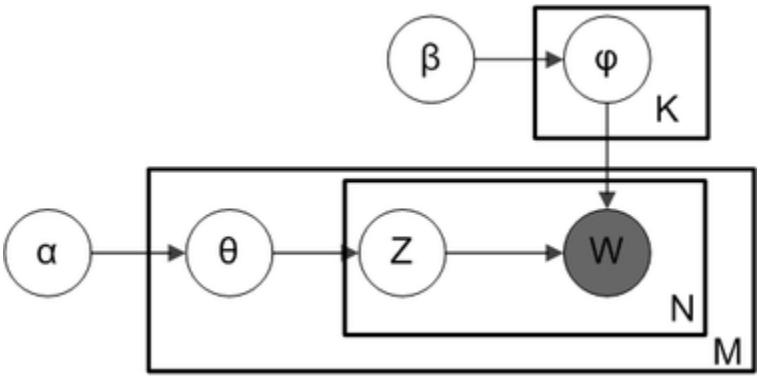

Figure 9: Plate notation for LDA

Titov and McDonald (2008a) showed that it is not effective to extract aspects using general LDA because it is based in word co-occurrence and topic differences and a set of reviews about a particular product almost always talk about the same aspects, which make the documents homogenous. They argued that it is useful in extracting entities in the reviews such as product or brand names. In their work, they developed a multi-grain topic model where they used global topic model to discover entities and a local model that finds aspect considering documents as a sliding window over a review. Most of the works followed were extension of LDA and were actually a joint model of aspect and sentiment.

While topic models (pLSA and LDA) can be modified to extract aspects alone (Mei et al., 2007), most of the models in the literatures were extensions of LDA and they model both aspect and sentiment (Jo & Oh, 2011; C. Lin & He, 2009; Titov & McDonald, 2008b; Zhao, Jiang, Yan,

& Li, 2010). There have been other models that predict aspect rating for the aspect extracted (Titov & McDonald, 2008a). We focus on unsupervised topic models that extract aspects along with sentiments, so we selected models that extract both aspects and sentiments. We have also added general LDA model described in this section as a baseline. The other three models are described in the following sections.

5.7.2 Aspect Sentiment Unification Model (ASUM)

ASUM was developed by Jo and Oh (2011) as an extension to their Sentence-LDA model which discovers aspects. ASUM aims at discovering aspects along with their sentiments. The generative model of ASUM is as follow

1. For every pair of sentiment s and aspect z, draw a word distribution $\phi_{sz} \sim Dirichlet(\beta_s)$
2. For each document d,
    a. Draw the document's sentiment distribution $\pi_d \sim Dirchilet(\gamma)$
    b. For each sentiment s, draw an aspect distribution $\theta_{ds} \sim Dirchlet(\alpha)$
    c. For each sentence,
        i. Chose a sentiment $j \sim Multinomial\ (\pi_d)$
        ii. Given sentiment j, chose an aspect $k \sim Multinomial(\theta_{dj})$
        iii. Generate words $w \sim Multinomial(\phi_{jk})$

The plate notation for ASUM is shown in Figure 10

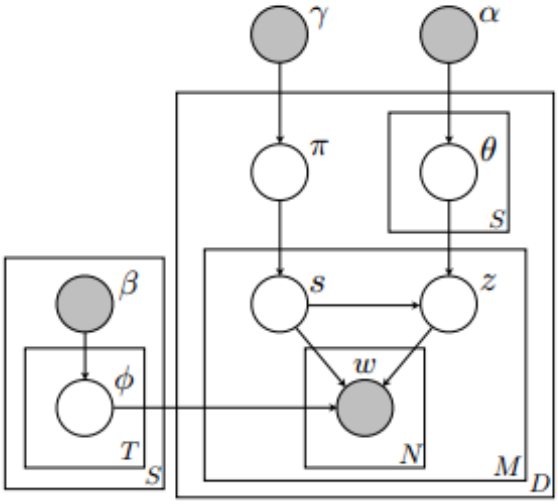

Figure 10 Plate Notation for ASUM

5.7.3 Joint Sentiment Topic Model (JST)

JST was developed by C. Lin and He (2009). JST is very similar to ASUM. They both model sentiment along with aspects. Unlike JST, ASUM limits individual words in a sentence to come from the same language model. This property captures the regional co-occurrence of a word in a document which leads to a model that is more focused. Furthermore, while both models rely on a seed sentiment word, JST makes use of them implicitly to allow the model to differentiate between positive and negative sentiments while ASUM integrates the set in the generative process which provide a stable statistical foundation for the model.

The generative process of the model as follow

1. For each document d, choose a distribution $\pi_d \sim Dirchlet(\gamma)$
2. For each sentiment label l under document d, choose a distribution $\theta_{d,l} \sim Dirchlet(\alpha)$.
3. For each word ω_i in document d
    a. Choose a sentiment label $l_i \sim \pi_d$,
    b. Choose a topic $z_i \sim \theta_{d,l_i}$,
    c. Choose a word w_i from the distribution over words defined by the topic z_i and sentiment label l_i, $\varphi_{z_i}^{l_i}$

The plate notation for JST is shown in Figure 11.

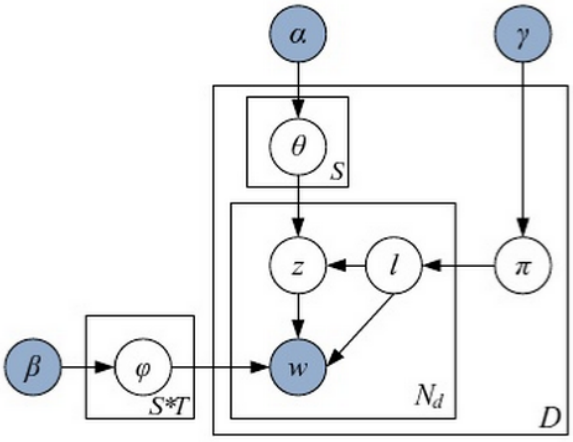

Figure 11: Plate Notation for JST Model

5.7.4 MaxEnt-LDA

This Model was developed by (Zhao et al., 2010). MaxEnt -LDA first captures general opinion words and then captures aspect specific words. The model works as follow:

1. Draw a background word distribution $\phi^B \sim Dirchlet(\beta)$

2. Draw a general aspect word distribution $\phi^{A,g} \sim Dirchlet(\beta)$
3. Draw a general opinion word distribution $\phi^{O,g} \sim Dirchlet(\beta)$
4. Draw a specific (0) and generic (1) type distribution $p \sim Beta(\gamma)$
5. For each aspect $t \in \{1, \dots, T\}\ do$
    a. Draw an aspect word distribution for aspect $t, \phi^{A,t} \sim Dirchlet(\beta)$
    b. Draw an aspect- specific opinion word distribution for aspect $t, \phi^{O,t} \sim Dirchlet(\beta)$
6. For each document $d \in \{1, \dots, D\}\ do$
    a. Draw an aspect distribution for document $d, \theta^{d} \sim Dirchlet(\alpha)$
    b. For each sentence $s \in \{1, \dots, S_d\}$ do
        i. Draw an aspect assignment $z_{d,s} \sim Multinomial(\theta^{d})$
        ii. For each word $w_{d,s,n}$ in sentence $s, n\ \in \{1, \dots, N_{d,s}\}$ do
            1. Set a background (0), aspect (1), and opinion (2) type distribution $\pi_{d,s,n} \leftarrow MaxEnt(x_{d,s,n}, \lambda)$
            2. Draw an assignment for indicator $y_{d,s,n} \sim Multinomial(\pi_{d,s,n})$
            3. Draw an assignment for indicator $u_{d,s,n} \sim Bernoulli(p)$
            4. Draw $w_{d,s,n} \sim \begin{Bmatrix} Multinomial(\phi^{B})\ if\ y_{d,s,n} = 0 \\ Multinomial(\phi^{A,z_{d,s}})\ if\ y_{d,s,n} = 1, u_{d,s,n} = 0 \\ Multinomial(\phi^{A,g})\ if\ y_{d,s,n} = 1, u_{d,s,n} = 1 \\ Multinomial(\phi^{O,z_{d,s}})\ if\ y_{d,s,n} = 2, u_{d,s,n} = 0 \\ Multinomial(\phi^{O,g})\ if\ y_{d,s,n} = 2, u_{d,s,n} = 1 \end{Bmatrix}$

The plate notation for MaxEnt-LDA is shown in Figure 12.

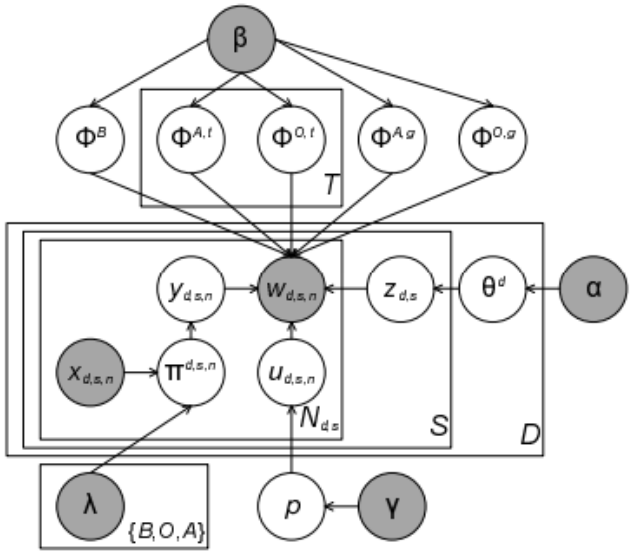

Figure 12: Plate Notation for MaxEnt-LDA

5.7.4.1 Model Seed Words: ASUM and JST both rely on a set of sentiment words used as a seed in the model. In their original work C. Lin and He (2009) use the PARADIGM set for JST. It consists of a set of positive and negative words that are used to define positive and negative sentiment orientation. The list is derived from the work of Pang, Lee, and Vaithyanathan (2002) for their baseline results. Their list is shown in Figure 13. Jo and Oh (2011) in their ASUM model used two sets of seed words PARADIGM (Bold) and PARADIGM+(all). The list is derived from Turney and Littman (2003) work. The list is shown in Figure 14.

Positive	Dazzling, brilliant, phenomenal excellent fanatic gripping mesmerizing riveting spectacular cool awesome thrilling moving exciting love wonderful best great superb still beautiful
Negative	Sucks terrible awful unwatchable hideous bad clichéd boring stupid slow worst waste unexcit rubbish tedious unbearable pointless cheesy frustrated awkward disappointing

Figure 13 Paradigm set used by (C. Lin & He, 2009)

Positive	**Good, Nice, excellent, positive, fortunate, correct, superior**, amazing, attractive, awesome, best, comfortable, enjoy, fantastic, favorite, fun, glad, great, happy, impressive, love, perfect, recommend, satisfied, thank, worth.
Negative	**Bad, nasty, poor, negative, unfortunate, wrong, inferior**, annoying, complain, disappointed, hate, junk, mess, not_good, not_like, not_recommend, not_worth, problem, regret, sorry, terrible, unacceptable, upset, waste, worst, worthless.

Figure 14 Full list of sentiment words PARADIGM and PARADIGM+.

5.7.5 Aspect Extraction Using Topic Modeling Applied to Arabic Dialect

We used these three models (ASUM, JST and MaxEnt-LDA) described in previous sections and we applied them to Arabic and translated reviews. We also used the original LDA model (described in section 5.7.1) as a base line. We experimented with these models with and without translation.

For the seed words for the models we used the same set in Figure 13 and Figure 14. In the case of Arabic reviews, we used word-to-word translation of the same set.

5.8 EVALUATION

5.8.1 Evaluating Extracted Aspects

Table 8. SEMEVAL2016 Dataset

	# review	#sentences	aspects	#distinct aspects
SemEval	1839	4802	9760	1616
RR	500	-	2263	1000

In order to evaluate the performance of these methods for the task of aspect extraction, we applied the methods to our datasets RR and TRR. We also utilized Semantic Evaluation workshop (SemEval 2016) dataset that was built for the task of aspect extraction (Pontiki et al., 2016). More specifically, the data set has been prepared to support the Arabic track of Task5: Aspect Based Sentiment Analysis which was part of Semantic Evaluation Workshop 2016 (SemEval 2016). Unfortunately, the competition received no submissions for this task for Arabic and therefore, there are no result that we can compare the performance against but they do provide the baseline for this dataset as F-score of 30.978 (Pontiki et al., 2016). The dataset contains hotel reviews

collected from Booking.com. We used the training dataset for subtask 2 which is aspect extraction. Statistics are provided in

Table 8 along with statistics for RR and TRR.

Under each method, we compared the performance of the method using precision, recall and F-score. 10-fold cross validation was used for the methods that require training and testing to overcome the problem of the limited amount of labeled data we have.

For our task, precision is defined as the ratio of the aspect retrieved to the number of aspect and non-aspect term retrieved. (How many of the returned aspects were correct). Similarly, recall is the ratio of the aspect term retrieved to the total aspect terms in the reviews (How many of the correct aspects were returned). F-score is the harmonic mean of precision and recall and is defined using Equation 1. F-score is used to evaluate the performance of different models and to compare systems. For our task, f-score was used to compare the four methods and the various modifications among each.

$$F = 2\,\frac{Precision\,.\,Recall}{Precision + Recall}$$

Equation 1. *F-score*

The goal of aspect extraction is to identify aspects that have been discussed in the review. Some of the topic models (ASUM and JST) we are using for extracting aspects output aspects along with their sentiments. Hence, we needed a way to separate opinions from aspects to be able to evaluate aspects alone, which is the focus of this dissertation. We separated aspects from opinions using two functions. The first function assumes that each word in the output is an aspect if it does not appear in our annotated corpus of opinion (description can be found in section 4.1.3.2). the second function follows the common notion that all opinion words are adjectives. After the separation, we

can follow the same path for evaluation as for the other methods by calculating precision, recall and f-score.

5.8.2 Evaluating Extracted Opinion Words

Double propagation, frequent nouns and noun phrases and topic models extract opinion words along with aspects. Therefore, we need to evaluate those methods and compare their performance. We also used precision, recall and f-score for this evaluation. In the case of topic models (ASUM and JST) that output aspect and opinions together, we used the two separation functions that are described in the previous section to able to calculate precision, recall and f-score

5.9 SUMMARY OF RESEARCH METHOD

Method	Type	Tools	Output		Evaluation
			Aspects	Opinions	
CRF	Supervised	POS Parser NER	YES	NO	Precision Recall Accuracy
Double propagation	Semi-supervised	POS Parser	YES	YES	Precision Recall
Frequent N and NP	Unsupervised	POS Association mining	YES	YES	Precision Recall
Topic Models	Unsupervised	-	YES	YES	Precision Recall

6.0 RESULTS

6.1 CRF RESULTS

We applied 10-fold cross validation CRF to the three datasets using a combination of the features discussed in section 5.4.2.1. We extracted the combination of features to be used in each model using Stanford Core NLP suite for translated reviews (TRR). We also examined the performance of this approach on Dialectal Arabic reviews in RR and SemEval. In this case, we extracted the features using MADAMIRA and Camel dependency parser. Table 9 shows the result of building different CRF models using various features in terms of F-scores. For simplicity Table 9 shows the F-score of those models and Appendix B shows the precision and recall of those models. Note that sentiment words (sw) and word distance (wd) cannot be applied to SemEval dataset because of the labeling used in that dataset and that the original task that the dataset was used for did not require identifying sentiment words for each aspect (Pontiki et al., 2016). Similarly, super part of speech tags (spos) was created by our annotators for RR dataset only.

Table 9. F-scores for CRF models on TRR, RR and SemEval Datasets
Read vertically, green shows low values and blue high values. Pink shows POS and SPOS impacts

Model Level	Features	Dataset		
		TRR	RR	SemEval
		F-score	F-score	F-score
Single	pos	0.834	0.764	0.7915
	ne	0.827	0.662	0.7920
	sp	0.833	0.751	0.7882
	wd	0.829	0.736	-
	sw	0.844	0.711	-
	spos	-	0.762	-
	Mean	**0.8334**	**0.731**	**0.7906**
Double	pos,ne	0.831	0.764	0.7901
	pos,spos	-	0.791	-
	pos,sp	0.831	0.785	0.7919
	pos,wd	0.833	0.810	-
	pos,sw	0.854	0.755	-
	ne,spos	-	0.764	-
	ne,sp	0.831	0.748	0.7873
	ne,wd	0.827	0.766	-
	ne,sw	0.844	0.713	-
	spos,sw	-	0.773	-
	spos,sp	-	0.782	-
	spos,wd	-	0.809	-
	sp,wd	0.829	0.765	-
	sp,sw	0.844	0.758	-
	wd,sw	0.840	0.776	-
	Mean	**0.8364**	**0.7527**	**0.7898**
Triple	pos, ne, sp	0.831	0.786	0.7926
	pos,ne,wd	0.831	0.810	-
	pos,ne,sw	0.852	0.776	-
	pos, sp,wd	0.833	0.812	-
	pos,sp,sw	0.858	0.792	-
	pos,sw,wd	0.858	0.808	-
	pos, spos,sw	-	0.807	-
	pos,spos,ne	-	0.790	-
	pos,spos,sp	-	0.805	-
	pos,spos,wd	-	0.825	-
	spos,sw,ne	-	0.773	-
	spos,sw,sp	-	0.787	-
	spos,sw,wd	-	0.803	-
	spos,ne,sp	-	0.785	-

	spos,ne,wd	-	0.811	-
	spos,sp,wd	-	0.808	-
	ne,sp,wd	0.827	0.766	-
	ne,sp,sw	0.842	0.751	-
	ne,wd,sw	0.838	0.744	-
	sp,wd,sw	0.840	0.776	-
	Mean	**0.841**	**0.7908**	**0.7926**
4+	**pos,ne,sp,wd**	0.833	0.810	-
	pos,ne,sp,sw	0.856	0.795	-
	pos,ne,wd,sw	0.858	0.807	-
	pos,sp,wd,sw	0.858	0.808	-
	pos,spos,sw,wd	-	0.817	-
	pos,spos,sw,ne	-	0.802	-
	pos,spos,sw,sp	-	0.8116	-
	pos,spos.ne.sp	-	0.806	-
	pos,spos,ne,wd	-	0.827	-
	pos,spos,sp,wd	-	0.824	-
	spos,sw,ne,sp	-	0.782	-
	spos,sw,ne,wd	-	0.805	-
	spos,sw,sp,wd	-	0.796	-
	spos,ne,sp,wd	-	0.811	-
	ne,sp,wd,sw	0.858	0.774	-
	pos,spos,sw,ne,sp	-	0.814	-
	pos,spos,sw,ne,wd	-	0.819	-
	pos,spos,sw,sp.wd	-	0.828	-
	pos,spos,ne,sp,wd	-	0.828	-
	pos,sw,ne,sp,wd	0.856	0.809	-
	spos,sw,ne.sp.wd	-	0.798	-
	pos,spos,ne,sp,wd,sw	-	0.819	-
	Mean	**0.8531**	**0.7965**	**-**
Overall Mean		**0.8408**	**0.778**	**0.7905**

We examined how supervised CRF models performed on the datasets. The results are shown in Table 9. In general, CRF models performed at a comparable level on SemEval and RR datasets (average F-score 0.7905 vs 0.778). This performance was better than the baseline provided by SemEval-2016 workshop which shows an F-score of 0.3152 (Pontiki et al., 2016). Furthermore, translated dataset (TRR) has the highest average with F-score of 0.8408. Appendix B shows the highest performing models in terms of precision and recall. Note that the highest performing models are highlighted in blue shades, the lowest performance is highlighted in green shades and the impact of POS and SPOS in pink shade in Table 9. We can see that translation performed well for translated Arabic dialect along with CRF for aspect extraction compared to datasets in Arabic Dialects. We relate the higher performance of translated reviews TRR in CRF to more accurate feature extraction because of the use of Stanford NLP tools. The main difference in developing CRF models in translated reviews TRR and Arabic dialect reviews RR and SemEval is the tools used to build the features used in the models. In the case of English translated reviews, we used Stanford Core NLP tool which has a reliable and more accurate performance. In the case of the Arabic reviews, we used MADAMIERA and Camel for dependency parsing. Those tools were not specifically developed for Arabic dialects and in the case of RR, Gulf dialect. In general, the lack of dialect tools contributed to the lower performance of the method. Also, the translated reviews have a better sentence structure than the original sentence structure. This observation was also highlighted by Refaee and Rieser (2015) in their research they related the better performance of the translated reviews using Microsoft Being to its ability to output a better sentence structure.

Examining the results more closely, we found that the best performing CRF models (highlighted in blue) in TRR contains sentiment words (sw) as a feature. Also, all of them except one contain word distance (wd) as a feature. Furthermore, all of them are combination of features

extracted using NLP tool except sentiment word which is human annotated opinion words that are translated from RR. Since the performance of CRF depends highly in the features extracted and most of those features extracted using NLP tools. This shows the important role that NLP plays in extracting those features. In the case of the translated review TRR, the features were extracted using Stanford NLP tools which was more accurate than the Arabic NLP tools we used in RR and SemEval.

On the other hand, RR and SemEval performed at a comparable level with SemEval performing better than RR. We contribute the better performance of SemEval than RR to the amount of the data. SemEval is a larger set than RR which provide a better training set. Also, SemEval reviews are longer than RR which provide a better training and probably provide a better probability of features per sentence.

We also found that the second best performing model in SemEval is using named entity tags (NE) as a single feature and the same model is the lowest performing model in RR. Examining these models more closely, we found that the NER tagger we used MADAMIRA identified more NER labels for SemEval than RR. Table 10 shows the distribution of those labels and their types. By manually examining the tags in both datasets we noticed that SemEval hotel reviews contains a lot of mention of different city names. In the case of RR, the dataset did not have a lot of city names and but it does have mention of some street names. In the case of Person tag (PER) a lot of the identified named entity tags were actually adjectives. Note that a lot of Arabic person names are adjective (for example the name 'Saeed سعيد' means happy and it does have the same Arabic spelling as the adjective 'happy سعيد '). The main point of this is that RR contains less named entities which led to both RR and TRR have the lowest performing models (green shade) when using named entity as a single feature (NE). The identification of named entity was performed

using Stanford Core NLP Named entity tagger for TRR and using MADAMIRA for RR. It also caused SemEval to have a better performing model when using named entity as a single or combined feature (blue shade).

Table 10. NER tags distribution on RR and SemEval

	RR	**SemEval**
Location	175	1687
Organization	42	539
Person	247	1410
O	9849	97047
Total tokens	10313	100683

The best performing models for RR contain five features which are (pos, spos, sw, sp, wd and pos, spos, ne, sp, wd). On the other hand, the best performing model for SemEval is the model that uses named entity as a single feature and the model that uses (pos,ne,sp). In general, the single and double feature models in SemEval performed better than RR (single model average 0.7906 vs 0.731double model average 0.7898 vs 0.7527) and the triple model performed comparably (0.7926 vs 0.7908). We hypothesis that adding more features to the model extracted from SemEval will slightly improve the performance. Unfortunately, we could not verify this hypothesis because of the lack of labels in SemEval. We would like to visit this in the future by adding the labels to SemEval.

When testing the models, we started with an individual feature and increased the features as we kept running the models. While single featured models performed somewhat similarly, there

was a slight improvement as we added more features in some combinations. Some of these combinations are shown in Table 11.

Table 11. Increased Precision and Recall as More Features are added

Dataset	Order	Combination	F-score
TRR	↓	**sw**	0.844
		sw,pos	0.854
RR	↓	**ne**	0.694
		ne,spos	0.764
		pos,spos,ne	0.790
RR	↓	**pos**	0.764
		pos,spos	0.791
		pos,spos,wd	0.825
		pos,spos,wd.sp	0.824

We also note that our experiment with Super part of speech (SPOS) to support POS tags for Arabic reviews had a slight improvement over using POS alone. CRF model with POS alone has a f-score of 0.764 and SPOS alone has f-score of 0.762 but CRF model with POS and SPOS together gives f-score of 0.791 (numbers are highlighted in red shades in Table 9). This slight improvement is attributed to a higher recall. For simplicity, the precision and recall scores are listed in Appendix B. We believe that the use of SPOS led to a better accuracy by supporting POS tags produced by NLP tagger.

6.2 DOUBLE PROPAGATION RESULTS

We examined the performance of double propagation aspect extraction method on DA reviews (RR and SemEval datasets) and on translated reviews. Similar to CRF feature building, we used MADAMIRA and CAMEL as the NLP tools required to run the algorithm. Furthermore, we also applied the method to TRR dataset to see how the method performs with translated text. We incorporated Stanford NLP suit along with Minipar as explained in section 5.5.2. We also examined the performance of this method on extracting sentiment words associated with each aspect. The results are illustrated on Table 12. Note that extracting sentiment words cannot be evaluated for SemEval dataset because of the lack of labeling.

Table 12. Double Propagation Results

Output	Dataset	Precision	Recall	F-score
Aspect	TRR	0.378	0.326	0.350
Aspect	RR	0.271	0.243	0.256
Aspect	SemEval	0.1824	0.563	0.276
Sentiment	TRR	0.5	0.829	0.624
Sentiment	RR	0.5	0.518	0.509

Table 12 shows that semi-supervised double propagation did not perform as well as supervised CRF models. This method exhibited low precision and recall in extracting aspect on the three datasets. Double propagation on Arabic datasets (RR and SemEval) performed lower than SemEval baseline (F-score 31%) for aspect extraction. On the other hand, this method performed slightly better on translated dataset (TRR) (35%).

These low results can be explained as follows. First of all, we have exhibited a loss of precision during translation that is attributed to the nature of Arabic dialect and the lack of state of the art Arabic Dialect translation tool. Furthermore, we also noticed that many times the sentence structure has a huge impact on the quality of translation. Unfortunately, Arabic Dialect does not have a clear sentence structure especially in social media text such as reviews which consequently affected the output of the translation.

Finally, double propagation relies on two things to succeed. The output of dependency parser and POS tagger and a set of rules that defines opinion, aspects and their relationship. Since the quality of the translation was affected, the output of the dependency parser was affected as well, which contributed to the low precision.

In the case of the Arabic reviews datasets (RR and SemEval), we observed that the nature of Arabic dialect and the lack of adequate NLP tools designed specifically for Arabic dialect affected the results. The output of camel dependency parser was not very accurate because it was developed for MSA and not Arabic dialect. Similarly, MADAMIRA was not specifically designed to handle Gulf Arabic dialect. MADAMIRA did perform well in the case of CRF which unlike double propagation does not rely on sentence structure. For example, finding the aspect “food” in a sentence like “the restaurant has a good food” depends on successfully labeling food as noun and correctly identifying the dependency relation “mod” between good and food.

Finally, we consider the fact that the rules were specifically designed for English reviews, affected the results in the case of Arabic review and it does also explain why the translated reviews performed better. The lack of NLP tools designed for Arabic dialect was a large factor in the performance of double propagation. Furthermore, we used MiniPar as a dependency parser for translated reviews and CAMEL for the Arabic reviews. Although both tools have a coarse tag set for part of speech, they are not the same. This may have played a role as well in the different performance between Arabic and translated reviews

6.3 FREQUENT NOUN AND NOUN PHRASES RESULTS

We examined how the frequent noun and noun phrases method performs on Arabic reviews (RR and SemEval dataset) and on translated reviews (TRR). The method also extracts all adjectives and consider them sentiment words, hence, we also evaluated the performance of this method on extracting sentiment words. The results are illustrated on

Table 13. Note that while the original method used stemming and considering the lack of accurate Arabic Dialect Stemmer we experimented with Arabic reviews with and without stemming. Similar to double propagation method, we could not assess the performance of this method for extracting sentiment on SemEval because of the lack of suitable labeling.

Table 13. Frequent noun and noun phrases results

Dataset	**Stemming**	**Output**	**Precision**	**Recall**	**F-score**
TRR	YES	Aspect	0.4	0.063	0.109
		Opinion	0.032	0.33	0.05
RR	NO	Aspect	0.51	0.102	0.169
		Opinion	0.134	0.39	0.2
	YES	Aspect	0.27	0.045	0.077
		Opinion	0.007	0.051	0.012
SemEval	NO	Aspect	0.6346	0.016	0.03
	YES	Aspect	0.7115	0.039	0.04

We examined the performance of this method on translated and Arabic reviews. We also examined the effect of stemming on precision and recall. We found that this method can achieve a reasonable precision on all three datasets but a very low recall. The SemEval dataset has the highest precision compared to TRR and RR as shown in

Table 13 in blue shades. This method performed better than double propagation but much lower than CRF. We were able to achieve 0.6346 and 0.7115 precision with SemEval dataset with and without stemming. The recall results for this method were significantly low for all datasets. Furthermore, because of the lack of state of the art Arabic dialect stemmer, we experimented with how the stemming affect the results in the case of Arabic reviews. Stemming lowered the precision with RR reviews but did improve the precision with SemEval dataset.

The low performance of this method is attributed to the algorithm itself and the nature of the reviews. First, the method performs association mining on all noun and noun phrases to find the most frequent aspects. The output of the association mining contains only 70 aspects in TRR

and about the same in RR and SemEval. We tried to play with minimum support to find a better output, put the best performing output was at 90% confidence. The method then takes those 70 frequent aspects and further performs pruning which reduces the total number of aspects. Moreover, the method then tries to extract infrequent aspects by identifying the closest adjectives to the identified frequent aspects and considers them opinion words. Finding other occurrences for these opinion words, it looks for the closest N and NP to find the infrequent aspects. We found that the initial low output of association rule contributed to the low results because there were not many opinion words to start with to look for infrequent feature and consequently increase the number of found aspects. Experimenting with larger dataset may improve the output of this method because a larger dataset will lead to more frequent aspect co-occurring which will give a larger set in the association mining step.

6.4 TOPIC MODEL RESULTS

We examined how the original topic model method LDA performed on our three datasets and we also experimented with variation of those models that were specifically designed for the task of aspect extraction. Note that we needed a function (F(OP)) to separate the opinions from aspects in the models that output them together. For that, we experimented with using our annotated opinions as a way to separate them and we also experimented with the notion that all opinions are mostly adjectives. We also would like to note that we could not apply JST and ASUM to SemEval dataset because of the lack of sentiment labeling. The results are listed in Table 14.

Table 14. Topic models results

Topic Model	Dataset	F(OP)	Output	Precision	Recall	F-score
ASUM	**TRR**	Annotated Opinion	Aspect	0.66	0.66	0.66
			Opinion	1	0.5	0.67
		All Adjectives	Aspect	0.685	0.685	0.685
			Opinion	0.18	0.205	0.198
	RR	Annotated Opinion	Aspect	0.34	0.34	0.34
			Opinion	1	0.7	0.82
		All Adjectives	Aspect	0.29	0.29	0.29
			Opinion	0.12	0.3	0.171
JST	**TRR**	Annotated Opinion	Aspect	0.565	0.565	0.565
			Opinion	1	0.46	0.63
		All Adjectives	Aspect	0.525	0.525	0.525
			Opinion	0.41	0.24	0.302
	RR	Annotated Opinion	Aspect	0.63	0.63	0.63
			Opinion	1	0.515	0.6798
		All Adjectives	Aspect	0.575	0.575	0.575
			Opinion	0.59	0.27	0.371
MaxEnt LDA	**TRR**	-	Aspect	0.27	0.27	0.27
			Opinion	0.102	0.102	0.102
	RR	-	Aspect	0.26	0.26	0.26
			Opinion	0.17	0.17	0.17
	SemEval	**-**	Aspect	0.174	0.174	0.174
LDA	**TRR**	-	Aspect	0.248	0.248	0.248
	RR	-	Aspect	0.306	0.306	0.306
	SemEval	-	Aspect	0.398	0.398	0.398

In general, topic models exhibited better performance than double propagation but a lower performance than CRF. We would like to discuss those results by comparing the performance of each model on the three datasets and discuss the implications. We then compare those models to each other. Note that the implementation of ASUM and JST output aspect and opinions together and we had to separate them to calculate precision and recall in a way similar to the other methods used. The methods we used are explained in section 5.8.1. To facilitate comparing the results between the data sets, the blue, green and yellow shades represent aspect extraction on TRR, RR and SemEval, respectively.

The first model is ASUM. ASUM model performed better with translated reviews TRR and much lower with Arabic reviews RR (blue vs green shade). The main difference in applying the model to both sets is the use of the seed words. The translation produces an Arabic MSA word and not Arabic dialect which may or may not appear in the reviews. We believe that this fact played a significant role in the performance of the model. We also suspect that our use of machine translation of the seeds words had a negative effect on the performance of ASUM on RR.

Similarly, JST performed better on RR and on a similar level to its performance on TRR (blue vs green shade) because JST does not incorporate the seed words in the model but uses it as a guide to separate negative and positive opinion. Evaluating opinion polarity is beyond the scope of this dissertation and thus will not be evaluated.

MaxEnt LDA model performed lower than JST and ASUM and even lower than semantic evaluation workshop baseline (F-score of 0.3152) (Pontiki et al., 2016). We attributed the low performance to the lack of the use of seed words and the fact that MaxEnt LDA relies on part of speech tags which are not very accurate in the case of Arabic dialect datasets (SemEval and RR).

Unlike the previous models which were implemented mainly to capture aspects and their sentiment, LDA is considered the base model from which the others were depicted. In line with that, LDA performance is lower than the other models in terms of aspect identification. Some of the previous work suggested that LDA is good for capturing global aspects which will be repeated throughout the reviews. We manually examined the output and found that many of the aspects that were outputted by LDA are global aspects such as (order, taste, desert, restaurant, service, room, location etc.).

7.0 DISCUSSION

7.1 ASPECT EXTRACTION METHODS PERFORMANCE

We examined the task of extracting aspect for sentiment analysis for Arabic dialectal reviews. We applied the most popular methods that have been used for English reviews. We also considered the limited availability of Arabic NLP as well as utilizing machine translations as a means to employ English NLP tools. We compared the performance of those four methods on three datasets: RR, TRR and SemEval. We also compared the performance between those four methods.

We found out that the best performing method was supervised CRF which performed better than all other methods on all datasets. We also found out that the other methods did not work and performed much lower than CRF. The main reason for the success of CRF is that it is a supervised method and relies on training and a set of features that are considered the basis for the model. Most of the features that we used are simple and can be obtained using simple NLP tools thus making CRF suitable for low resource languages such as Arabic. The main drawback of this method is that it requires a dataset with labels and the larger the dataset the better the model performance as is the case with most supervised methods. The availability of a dataset specifically labeled for this task might be a challenge for many low resource languages. Furthermore, CRF on translated reviews did better than Arabic dialect reviews RR and SemEval. This success is comparable to the success of CRF on English aspect extraction. Refer to Jakob and Gurevych (2010) for comparable

results in single and multi-domain aspect extraction. The better performance of CRF in translated reviews is because translating the reviews from Arabic dialect to English provided us with a way to overcome of the low performance of Arabic NLP tools on RR and SemEval. Additionally, the sentences produced by the translators exhibit a better sentence structure than the original reviews which also contributed to better NLP performance. We note that the performance of RR and SemEval was reasonable and comparable to single domain English aspect extraction in Jakob and Gurevych (2010). We suspect that the performance of CRF on RR and SemEval can be improved by using NLP tools specifically designed for Arabic dialects.

Although there was no major difference between varying the features of CRF, we believe that some of these features are promising and should be experimented with more to prove their importance. Sentiment words (sw) which are labels identified by our annotators achieved some of the highest scores in both RR and TRR. Regrettably, we could not experiment with those labels on SemEval because of the lack of annotations. The other promising features were SPOS and Named entity. Although our datasets TRR and RR did not contain as much identified NE as SemEval we believe that named entities are domain dependents and in the case of hotel reviews in SemEval, users tend to mention more name entities than in the restaurant domain. This is not to generalize this as a fact but to raise this difference and the role it played in our results. Lastly, shortest dependency path (SP) and word distance (WD) can be improved by using NLP tools designed for Arabic dialects. The use of dependency parser designed for Arabic MSA on our dialectal dataset contributed to the performance of those features.

Surprisingly, the unsupervised method topic models performed better than semi-supervised double propagation and unsupervised frequent noun and noun phrases. The impact of NLP tools has a huge factor in the performance of double propagation and frequent noun and noun phrases

which lead to a much lower performance compared to the other methods that do not employ as much NLP tools and do not depend on sentence structure. By manually tracing the application of double propagation rules on our RR reviews, we found that the rules failed to identify the correct label either because of the sentence structure or because of incorrect labeling produced by sentence dependency parser. We believe that double propagation can be improved by implementing a set of rules that relies on the sentence structure of Arabic dialect. We also believe that the use of a dependency parser that is specifically implemented for Arabic dialect will help identify aspects and opinion words better. Frequent noun and noun phrases method can be improved by using a larger dataset that will lead to more noun and noun phrases occurrences which will boost the output of association mining which is the first step of the algorithm that failed to identify aspects in our datasets. The identification of opinion words on frequent noun and noun phrases is totally dependent upon the initial step of finding frequent nouns by using association mining algorithm. We believe that this can be improved by using a larger dataset but we also believe that using opinion seed words instead of depending on extracted frequent aspects will improve the performance of finding infrequent aspects. Instead of assuming that the opinion word is the closest adjective to the identified aspect, using a good dependency parser may contribute to better performance by finding adjective that are associated with the aspect that are not necessarily the closest to the aspect.

While some of the topic models exhibit better performance, there is still room for improvement. The performance of ASUM and JST were better than all the other models we examined. We believe that the use of seed words had a huge impact in boosting the results. ASUM showed some promising results to succeed as aspect extraction but we need to use seed words drawn from Arabic Dialects and not translated version of the English list. Maxent LDA did not

show a superior performance to the other three models but we would like to experiment with it in the future by using a larger dataset and see whether the performance will improve.

While we did not see much improvement with using variety of features in training CRF models, it still performed better than the other approaches. Consequently, we recommend CRF for domain specific Arabic aspect extraction and any other method that utilizes POS without relying heavily on dependency parsing or sentence structure given the current state of NLP tools for Arabic dialects.

7.2 OPINION EXTRACTION METHODS PERFORMANCE

Since some of these methods extract opinion words along with aspects, we evaluated how well those methods performed on extracting opinion words in Arabic reviews with and without translation. Double propagation has 50% precision for both Arabic and translated reviews but it has a higher recall for translated reviews at 83% compared to 51% for Arabic. The reason for the higher performance for translated reviews is the use of seed words along with that the rules were originally crafted for English reviews and follow the English sentence structure.

Frequent noun and noun phrases method performed much lower because of the original performance of this method in extracting aspects. This method follows a series of steps that depend on each other. It extracts aspects using association mining and then extracts the closest adjective to those aspects as opinion words. Since, association mining had few aspects, there were few opinions extracted as well. Note, that Arabic reviews without stemming performed higher than translated reviews.

In the case of topic models and as discussed in the evaluation section 5.8.1, we had to deal with separating opinion and aspect words from the output of ASUM and JST. We used two methods: the first uses all annotated opinion in our dataset and the second uses the assumption that all adjectives are opinions. JST and ASUM performed better than MaxEnt-LDA. JST performed better than ASUM in extracting opinion for both Arabic and English in terms of precision. On the other hand, ASUM achieved better recall in all variations (highest recall 70%). We also noted that ASUM performed better with translated reviews while JST performed better with Arabic. We relate this difference in performance to the nature of each model and how it is implemented.

7.3 CONCLUSION

We examined how aspect extraction methods perform on Arabic Dialect reviews given the limitation of the Arabic dialect natural language processing field. We achieved a reasonable performance using supervised conditional random fields on an in-house developed dataset and SEMEVAL competition dataset. We also examined how other methods perform (double propagation, topic models and frequent noun and noun phrases) and how they can be applied without the need to develop extensive NLP resources. We examined how translation to a wealthy resource language, mainly English, can aid and overcome the shortcoming of NLP resources for Arabic and that translation can be used a mean to achieve aspect extraction for any low resource language. Out of the four examined methods supervised conditional random field gives the best performance on the three datasets. We also found that translation achieves higher results than using direct Arabic NLP tools. We also discussed why each method succeeded or failed and we included many suggestions on how each method can be improved. Given that there is some effort toward

building NLP tools for Arabic Dialects, it would be interesting to see how these tools will affect the performance of these methods. We also extended our contribution to cover low resource languages, they can benefit from applying the same approach to languages that suffer from the lack of NLP tools by implementing minor modifications.

7.4 RECOMMENDATION FOR FUTURE RESEARCH

The research on aspect extraction for sentiment analysis can be extended further by experimenting with a larger dataset which could further improve the performance of these methods. The development of such dataset requires time and resources that were not available to us at the time of this research but might be manageable in the future. We also would like to revisit these methods as NLP tools for Arabic dialects become available. Additionally, we plan to extend this field further by examining the methods for extracting aspect specific opinion, analyze their sentiment and categorize them. Finally, we would like to build a complete system that do sentiment analysis at the aspect level by identifying aspects, categorize them and classify their sentiment.

APPENDIX A

ARABIC LANGUAGE

Arabic belongs to the Semitic family of languages along with Hebrew, Aramaic, etc. It is the native language of 27 countries and there are 290 million native speakers. Arabic speakers are about 18.8% of the Internet users' population and it is one of the fastest growing populations on the web according to the Internet World Statistics Report.

The Arabic Language is divided into three types: Classical Arabic (CA), Modern Standard Arabic (MSA) and Dialect Arabic (DA). MSA is the only standard form that is widely recognized and formally taught in schools. MSA is based on Classical Arabic which is the language of the Qur'an (Muslims' holy book). The MSA is not a native language of any country and it is largely different from dialect forms. The Arabic language has many different dialects for every region that are used in informal daily communications. The dialects are not standardized or taught formally in schools.

Arabic shared vocabulary with other languages such as Farsi, Urdu and Malay. Arabic Alphabets consist of 28 letters (vowels and consonants). Table 15 shows the Arabic alphabet along with their English equivalents. Along with alphabets Arabic also has diacritics which are marks that are used as phonetic guide. Diacritics in Arabic are summarized in Table 16. While the use of

diacritics is optional in the written form of Arabic, it aids in resolving ambiguity between words that have the same letters. The use of diacritics in social media is not common.

Table 15. Arabic Alphabet

Arabic Letters	Key Word	Approximate Pronunciation	English Equivalent	IPA Symbol
أ	أب	Hamzah		ʔ
ا		Alif	à	æ
ب	بدر	Baa	B	b
ت	تمر	*Taa	T	t
ث	ثواب	*Thaa	Th as in “three”	ɵ
ج	جمل	Jemm	J	ʤ / g / ʒ
ح	حمل	Haa	-	ħ
خ	خال	Khaa	-	X
د	درس	*Daal	D	d
ذ	ذاكر	*Thaal	Th as in “then”	ð
ر	رزق	*Raa	R	r
ز	زمن	*Zaa	Z	z
س	سار	*Seen	S	s
ش	شمس	*Sheen	Sh	ʃ
ص	صبح	*Saad	-	s̲
ض	ضرب	Dhaad	-	d̲
ط	طب	*Tah	-	t̲
ظ	ظهر	*Dhaa	-	ð̲
ع	عين	*Ein	-	ʕ
غ	غرب	Ghein	-	ɣ
ف	فال	Faa	F	f
ق	قمر	Qaaf	-	q
ك	كتب	Kaaf	K	k
ل	لوح	*Laam	L	l
م	ماء	Meem	M	m
ن	نور	*Noon	N	n
هـ	هلال	Haa	H	h
و	وقف	Wow	W	w / u
ي	يقف	Yaa	Y	j / l

Note: Adapted from (Alorifi, 2008)

Table 16. Arabic Diacritics

◌َ	Fatha	Short /a/ sound
◌ُ	Dhamah	Short /u/ sound
◌ِ	Kasrah	Short /i/ sound
◌ْ	Sokon	Indicate that the constant is not followed by vowel
◌ٍ	Tanween Kasr	In sound at the end of the word
◌ً	Tanween Fath	a sound at the end of the word
◌ٌ	Tanween Dham	Un sound at the end of the word
◌ّ	Shaddah	gemination (consonant doubling or extra length)
~	Maddah	Placed mostly on top of Alif letter and present long /a/ sound

Arabic script is a cursive and is written from right to left similar to other languages such as Farsi, Kurdish, and Pashto. The letters take different shapes depending on their position on the word. Arabic is a morphologically rich language. Word formation in Arabic is highly

derivational and is based on roots variations. A comparison between Arabic and English word formation is provided in Figure 15. Word formation in Arabic with English equivalents.

He wrote	Kataba	Past verb masculine	كَتَبَ
She wrote	Katabat	Past verb feminine	كَتَبَتْ
Writer	Kateb	Noun-single	كَاتِبْ
Writers	Kotab	Noun-plural	كُتْاب
Book	Ketab	Noun-single	كِتَابْ
Books	Kotob	Noun-plural	كُتُبْ
He Is writing	Yaktob	Present continues verb masculine	يَكْتُبْ
She is writing	Taktob	Present continues verb feminine	تكتب
Library	Maktabah	Noun singular	مَكْتَبَة

Figure 15. Word formation in Arabic with English equivalents

Arabic has both definiteness and indefinite markers. The definite marker is ‘al-’ which is prefix that attached to the beginning of noun and adjectives for example: ‘al-ketab-الكتاب’ is a nominal definitive for the book. The indefinite marker is Tanween (see Table 16) which is a sound added to the end of nouns for example: Ketabun-كتابٌ is nominal indefinite for book. While the definite marker is always written, the indefinite marker is usually omitted.

Arabic sentences have two forms: nominal and verbal sentences. The nominal consists of two consecutive words: A subject followed by adjective. The verbal pattern has two forms: Subject-Verb-Object and Verb- Subject-Object. At the higher level: Arabic has three part of speech: Nominal, Verb and Particles.

While MSA has many defined rules and clear sentence structure, Arabic dialects exhibit a different behavior. The sound of letters varies from region to region (Figure 16 shows this variability). Furthermore, even some nouns and verbs differ between dialects (Figure 17). There are no defined rules on how Arabic dialect words and sentence are formed. Arabic dialects also exhibit spelling inconsistency (Figure 18).

Letter	**ث**	**ج**	**ذ**	**ق**
Gulf	Th	J/g/y	TH	G
Levantine	T	J	Th	A
Egyptian	T	G	Z	A

Figure 16. Variation of phonemes in Arabic dialect

English	**MSA**		**Gulf**		**Levantine**		**Egyptian**	
Delicious	لذيذ	Latheeth	حلو	Helo	طيب	Taeeb	لزيز	Lazzez
Table	طاولة	Tawelahh	طاولة	Tawelah	طاولة	Tawelah	طربيزه	Tarabeezah
How are you?	كيف حالك؟	Keef Halok	شلونك	shloonok	كيفك؟	Keefak	ازيك	Ezzaek
Very	جداً	Jeedan	مره	Marrah	كتير	Kteer	أوي	Awe

Figure 17. Word variation between MSA and different dialects

MSA	Arabic Dialect
ضوء	ضو ضوء ضوا
قلب	الب كلب قلب
ايضا	برضو بردو
مكتبة	مكتبه مكتبة

Figure 18. Spelling variation in Arabic dialect

APPENDIX B

PRECISION AND RECALL RESULTS FOR CRF METHOD

Table 17 shows the results of all the CRF models run on the three datasets. The combined scores of those results. Blue shades represent the highest precision and recall. The green represents the lowest highest precision and recall. The red represents a special observation explained in section 6.1.

Table 17. Precision (P) and Recall (R) for CRF models on TRR, RR and SemEval Datasets (Read vertically, green identifies low values and blue indicated high values. Pink shows POS and SPOS impacts)

Model Level	Features	Dataset					
		TRR		RR		SemEval	
		P	R	P	R	P	R
Single	**pos**	0.931	0.789	0.855	0.728	0.850	0.753
	ne	0.932	0.781	0.846	0.662	0.848	0.754
	sp	0.937	0.786	0.854	0.712	0.849	0.748
	wd	0.936	0.782	0.854	0.700	-	-
	sw	0.943	0.798	0.867	0.675	-	-
	spos	-	-	0.860	0.725	-	-
	Mean	**0.9358**	**0.7872**	**0.856**	**0.7003**	**0.8493**	**0.7518**
Double	**pos,ne**	0.936	0.784	0.849	0.729	0.8499	0.751
	pos,spos	-	-	0.860	0.757	-	-
	pos,sp	0.936	0.784	0.864	0.747	0.8400	0.758
	pos,wd	0.937	0.786	0.871	0.775	-	-
	pos,sw	0.949	0.807	0.868	0.737	-	-
	ne,spos	-	-	0.861	0.727	-	-
	ne,sp	0.936	0.784	0.853	0.709	0.8485	0.748
	ne,wd	0.831	0.779	0.861	0.726	-	-
	ne,sw	0.943	0.798	0.868	0.677	-	-
	spos,sw	-	-	0.856	0.736	-	-
	spos,sp	-	-	0.858	0.746	-	-
	spos,wd	-	-	0.855	0.780	-	-
	sp,wd	0.936	0.782	0.864	0.725	-	-
	sp,sw	0.943	0.798	0.858	0.719	-	-
	wd,sw	0.942	0.793	0.857	0.739	-	-
	Mean	**0.9289**	**0.7895**	**0.8409**	**0.718**	**0.846**	**0.752**
Triple	**pos, ne, sp**	0.936	0.784	0.862	0.749	0.8485	0.759
	pos,ne,wd	0.936	0.784	0.871	0.775	-	-
	pos,ne,sw	0.949	0.805	0.872	0.737	-	-
	pos, sp,wd	0.937	0.786	0.872	0.778	-	-
	pos,sp,sw	0.938	0.815	0.860	0.757	-	-
	pos,sw,wd	0.938	0.815	0.863	0.776	-	-
	pos, spos,sw	-	-	0.854	0.779	-	-
	pos,spos,ne	-	-	0.863	0.755	-	-
	pos,spos,sp	-	-	0.857	0.774	-	-
	pos,spos,wd	-	-	0.866	0.798	-	-
	spos,sw,ne	-	-	0.865	0.736	-	-
	spos,sw,sp	-	-	0.855	0.752	-	-
	spos,sw,wd	-	-	0.841	0.779	-	-

	spos,ne,sp	-	-	0.859	0.748	-	-
	spos,ne,wd	-	-	0.856	0.783	-	-
	spos,sp,wd	-	-	0.853	0.780	-	-
	ne,sp,wd	0.935	0.779	0.861	0.726		
	ne,sp,sw	0.943	0.795	0.854	0.712	-	-
	ne,wd,sw	0.942	0.791	0.856	0.737	-	-
	sp,wd,sw	0.942	0.791	0.857	0.739	-	-
	Mean	**0.9396**	**0.7945**	**0.8599**	**0.7585**	**0.8485**	**0.7587**
4+	**pos,ne,sp,wd**	0.937	0.786	0.871	0.775	-	-
	pos,ne,sp,sw	0.938	0.813	0.861	0.760	-	-
	pos,ne,wd,sw	0.938	0.815	0.859	0.777	-	-
	pos,sp,wd,sw	0.938	0.815	0.863	0.776	-	-
	pos,spos,sw,wd	-	-	0.850	0.794		
	pos,spos,sw,ne	-	-	0.858	0.772	-	-
	pos,spos,sw,sp	-	-	0.845	0.788	-	-
	pos,spos.ne.sp	-	-	0.862	0.773	-	-
	pos,spos,ne,wd	-	-	0.863	0.802	-	-
	pos,spos,sp,wd	-	-	0.868	0.796	-	-
	spos,sw,ne,sp	-	-	0.852	0.747	-	-
	spos,sw,ne,wd	-	-	0.842	0.781	-	-
	spos,sw,sp,wd	-	-	0.838	0.769	-	-
	spos,ne,sp,wd	-	-	0.856	0.783	-	-
	ne,sp,wd,sw	0.938	0.815	0.856	0.737	-	-
	pos,spos,sw,ne,sp	-	-	0.842	0.794	-	-
	pos,spos,sw,ne,wd	-	-	0.848	0.798	-	-
	pos,spos,sw,sp.wd	-	-	0.866	0.802	-	-
	pos,spos,ne,sp,wd	-	-	0.866	0.802	-	-
	pos,sw,ne,sp,wd	0.938	0.813	0.860	0.779	-	-
	spos,sw,ne.sp.wd	-	-	0.839	0.772	-	-
	pos,spos,ne,sp,wd,sw	-	-	0.847	0.799	-	-
	Mean	**0.9378**	**0.8095**	**0.8425**	**0.7690**	**-**	**-**
Overall Mean		**0.9352**	**0.7946**	**0.8492**	**0.7467**	**0.8478**	**0.7531**

BIBLIOGRAPHY

Abbasi, A., Chen, H., & Salem, A. (2008). Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums. *ACM Transactions on Information Systems (TOIS), 26*(3), 12.

Abdul-Mageed, M., & Diab, M. T. (2011). *Subjectivity and Sentiment Annotation of Modern Standard Arabic Newswire.* Paper presented at the Linguistic Annotation Workshop.

Abdul-Mageed, M., & Diab, M. T. (2012a). *AWATIF: A Multi-Genre Corpus for Modern Standard Arabic Subjectivity and Sentiment Analysis.* Paper presented at the LREC.

Abdul-Mageed, M., & Diab, M. T. (2012b). *Toward building a large-scale Arabic sentiment lexicon.* Paper presented at the Proceedings of the 6th International Global WordNet Conference.

Abdul-Mageed, M., & Diab, M. T. (2014). *Sana: A large scale multi-genre, multi-dialect lexicon for arabic subjectivity and sentiment analysis.* Paper presented at the Proceedings of the Language Resources and Evaluation Conference (LREC).

Abdul-Mageed, M., & Korayem, M. (2010). *Automatic identification of subjectivity in morphologically rich languages: the case of Arabic.* Paper presented at the Proceedings of the 1st workshop on computational approaches to subjectivity and sentiment analysis (WASSA), Lisbon.

Abdul-Mageed, M., Kübler, S., & Diab, M. T. (2012). *Samar: A system for subjectivity and sentiment analysis of arabic social media.* Paper presented at the Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis.

Ahmed, S., Pasquier, M., & Qadah, G. (2014). Key Issues in Conducting Sentiment Analysis on Arabic Social Media Text.

Al-Smadi, M., Qawasmeh, O., Talafha, B., & Quwaider, M. (2015). *Human annotated arabic dataset of book reviews for aspect based sentiment analysis.* Paper presented at the Future Internet of Things and Cloud (FiCloud), 2015 3rd International Conference on.

Al-Subaihin, A. A., Al-Khalifa, H. S., & Al-Salman, A. S. (2011). *A proposed sentiment analysis tool for modern arabic using human-based computing.* Paper presented at the Proceedings of the 13th International Conference on Information Integration and Web-based Applications and Services.

Albraheem, L., & Al-Khalifa, H. S. (2012). *Exploring the problems of sentiment analysis in informal Arabic*. Paper presented at the Proceedings of the 14th International Conference on Information Integration and Web-based Applications & Services, Bali, Indonesia.

Alorifi, F. S. (2008). *Automatic Identification of Arabic Dialects Using Hidden Markov Models*: ProQuest.

Blair-Goldensohn, S., Hannan, K., McDonald, R., Neylon, T., Reis, G. A., & Reynar, J. (2008). *Building a sentiment summarizer for local service reviews.* Paper presented at the WWW Workshop on NLP in the Information Explosion Era.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research, 3*, 993-1022.

Borgelt, C. (2012). Frequent item set mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2*(6), 437-456.

Brody, S., & Elhadad, N. (2010). *An unsupervised aspect-sentiment model for online reviews.* Paper presented at the Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics.

Choi, Y., Cardie, C., Riloff, E., & Patwardhan, S. (2005). *Identifying sources of opinions with conditional random fields and extraction patterns.* Paper presented at the Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing.

El-Halees, A. (2011). *Arabic Opinion Mining Using Combined Classification Approach.* Paper presented at the Proceeding The International Arab Conference On Information Technology, Azrqa, Jordan.

Elarnaoty, M., AbdelRahman, S., & Fahmy, A. (2012). A Machine Learning Approach For Opinion Holder Extraction In Arabic Language. *arXiv preprint arXiv:1206.1011*.

Elhawary, M., & Elfeky, M. (2010). *Mining Arabic Business Reviews.* Paper presented at the Data Mining Workshops (ICDMW), 2010 IEEE International Conference on.

Farra, N., Challita, E., Assi, R. A., & Hajj, H. (2010). *Sentence-Level and Document-Level Sentiment Mining for Arabic Texts.* Paper presented at the Data Mining Workshops (ICDMW), 2010 IEEE International Conference on.

Habash, N. (2010). *Introduction to Arabic natural language processing* (Vol. 3).

Habash, N., & Roth, R. M. (2009). *Catib: The columbia arabic treebank.* Paper presented at the Proceedings of the ACL-IJCNLP 2009 conference short papers.

Hofmann, T. (1999). *Probabilistic latent semantic indexing.* Paper presented at the Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval.

Hu, M., & Liu, B. (2004a). *Mining and summarizing customer reviews.* Paper presented at the Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining.

Hu, M., & Liu, B. (2004b). *Mining opinion features in customer reviews.* Paper presented at the AAAI.

Jakob, N., & Gurevych, I. (2010). *Extracting opinion targets in a single-and cross-domain setting with conditional random fields.* Paper presented at the Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing.

Jin, W., Ho, H. H., & Srihari, R. K. (2009). *A novel lexicalized HMM-based learning framework for web opinion mining.* Paper presented at the Proceedings of the 26th Annual International Conference on Machine Learning.

Jo, Y., & Oh, A. H. (2011). *Aspect and sentiment unification model for online review analysis.* Paper presented at the Proceedings of the fourth ACM international conference on Web search and data mining.

Joshi, A., Balamurali, A., Bhattacharyya, P., & Mohanty, R. (2011). *C-Feel-It: a sentiment analyzer for micro-blogs*. Paper presented at the Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Systems Demonstrations.

Korayem, M., Crandall, D., & Abdul-Mageed, M. (2012). Subjectivity and Sentiment Analysis of Arabic: A Survey *Advanced Machine Learning Technologies and Applications* (pp. 128-139): Springer.

Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Li, F., Han, C., Huang, M., Zhu, X., Xia, Y.-J., Zhang, S., & Yu, H. (2010). *Structure-aware review mining and summarization.* Paper presented at the Proceedings of the 23rd international conference on computational linguistics.

Lin, C., & He, Y. (2009). *Joint sentiment/topic model for sentiment analysis.* Paper presented at the Proceedings of the 18th ACM conference on Information and knowledge management.

Lin, D. (2003). Dependency-based evaluation of MINIPAR *Treebanks* (pp. 317-329): Springer.

Liu, B. (2012). *Sentiment analysis and opinion mining* (Vol. 5).

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). *The Stanford CoreNLP natural language processing toolkit.* Paper presented at the Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations.

Marton, Y., Habash, N., & Rambow, O. (2010). *Improving Arabic dependency parsing with lexical and inflectional morphological features.* Paper presented at the Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages.

Marton, Y., Habash, N., & Rambow, O. (2013). Dependency parsing of Modern Standard Arabic with lexical and inflectional features. *Computational Linguistics, 39*(1), 161-194.

Mei, Q., Ling, X., Wondra, M., Su, H., & Zhai, C. (2007). *Topic sentiment mixture: modeling facets and opinions in weblogs.* Paper presented at the Proceedings of the 16th international conference on World Wide Web.

Mourad, A., & Darwish, K. (2013). Subjectivity and Sentiment Analysis of Modern Standard Arabic and Arabic Microblogs. *WASSA 2013*, 55.

Omar, N., Albared, M., Al-Shabi, A. Q., & Al-Moslmi, T. (2013). Ensemble of Classification Algorithms for Subjectivity and Sentiment Analysis of Arabic Customers' Reviews.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). *Thumbs up?: sentiment classification using machine learning techniques.* Paper presented at the Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10.

Pasha, A., Al-Badrashiny, M., Diab, M. T., El Kholy, A., Eskander, R., Habash, N., . . . Roth, R. (2014). *MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic.* Paper presented at the LREC.

Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., AL-Smadi, M., . . . Eryiit, G. (2016). SemEval-2016 task 5: Aspect based sentiment analysis. *In: Proceedings of the 10th International Workshop on Semantic Evaluation*.

Qiu, G., Liu, B., Bu, J., & Chen, C. (2011). Opinion word expansion and target extraction through double propagation. *Computational Linguistics, 37*(1), 9-27.

Rabiner, L., & Juang, B.-H. (1986). An introduction to hidden Markov models. *ASSP Magazine, IEEE, 3*(1), 4-16.

Refaee, E., & Rieser, V. (2015). *Benchmarking Machine Translated Sentiment Analysis for Arabic Tweets.* Paper presented at the NAACL-HLT 2015 Student Research Workshop (SRW).

Rushdi-Saleh, M., Martín-Valdivia, M. T., Ureña-López, L. A., & Perea-Ortega, J. M. (2011). Bilingual Experiments with an Arabic-English Corpus for Opinion Mining.

Salameh, M., Mohammad, S. M., & Kiritchenko, S. (2015). *Sentiment after translation: A case-study on arabic social media posts.* Paper presented at the Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

Soliman, T. H., Elmasry, M., Hedar, A., & Doss, M. (2014). Sentiment Analysis of Arabic Slang Comments on Facebook. *INTERNATIONAL JOURNAL OF COMPUTERS & TECHNOLOGY, 12*(5), 3470-3478.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology, 61*(12), 2544-2558.

Titov, I., & McDonald, R. (2008a). *A Joint Model of Text and Aspect Ratings for Sentiment Summarization.* Paper presented at the ACL.

Titov, I., & McDonald, R. (2008b). *Modeling online reviews with multi-grain topic models.* Paper presented at the Proceedings of the 17th international conference on World Wide Web.

Turney, P. D., & Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS), 21*(4), 315-346.

Zhao, W. X., Jiang, J., Yan, H., & Li, X. (2010). *Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid.* Paper presented at the Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing.

Zhuang, L., Jing, F., & Zhu, X.-Y. (2006). *Movie review mining and summarization.* Paper presented at the Proceedings of the 15th ACM international conference on Information and knowledge management.